Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent

Yue Li * Yuting Wei †

October 16, 2021

Abstract

An evolving line of machine learning works observe empirical evidence that suggests interpolating estimators — the ones that achieve zero training error — may not necessarily be harmful. This paper pursues theoretical understanding for an important type of interpolators: the minimum ℓ_1-norm interpolator, which is motivated by the observation that several learning algorithms favor low ℓ_1-norm solutions in the over-parameterized regime. Concretely, we consider the noisy sparse regression model under Gaussian design, focusing on linear sparsity and high-dimensional asymptotics (so that both the number of features and the sparsity level scale proportionally with the sample size).

We observe, and provide rigorous theoretical justification for, a curious *multi-descent* phenomenon; that is, the generalization risk of the minimum ℓ_1-norm interpolator undergoes multiple (and possibly more than two) phases of descent and ascent as one increases the model capacity. This phenomenon stems from the special structure of the minimum ℓ_1-norm interpolator as well as the delicate interplay between the over-parameterized ratio and the sparsity, thus unveiling a fundamental distinction in geometry from the minimum ℓ_2-norm interpolator. Our finding is built upon an exact characterization of the risk behavior, which is governed by a system of two non-linear equations with two unknowns.

Keywords: minimum norm interpolators, multiple descent, Lasso, sparse linear regression, exact asymptotics, approximate message passing

Contents

*Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
†Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

1 Introduction

At the core of statistical learning lies the problem of understanding the generalization performance (e.g., out-of-sample errors) of the learning algorithms in use. Conventional wisdom in statistics held that including too many covariates when training statistical models can hurt generalization (despite improving training accuracy), due to the undesired over-fit. This leads to the classical conclusion that: proper regularization — through either adding certain penalty functions to the loss function or algorithmic self-regularization — seems to be critical in achieving the desired accuracy (e.g., Friedman et al. (2001); Wei et al. (2019)). However, an evolving line of works in machine learning observes empirical evidence that suggests, to the surprise of many statisticians, over-parameterization is not necessarily harmful. Indeed, many machine learning models (such as random forests or deep neural networks) are trained until the training error vanishes to zero — meaning that they are able to perfectly interpolate the data — while still generalizing well (e.g., Zhang et al. (2021); Wyner et al. (2017); Belkin et al. (2019)). As a key observation to explain this phenomenon, many models when trained by gradient type methods (e.g., gradient descent, stochastic gradient descent, AdaBoost) converge to certain minimum norm interpolators, which implicitly favor models with smaller model complexity.

These empirical mysteries inspire a recent flurry of activity towards understanding the generalization properties of various interpolators. A dominant fraction of recent efforts, however, concentrated on studying certain minimum ℓ_2-norm interpolators, primarily in the context of linear and/or kernel regression (see, e.g., Liang and Rakhlin (2020); Mei and Montanari (2019); Hastie et al. (2019); Belkin et al. (2020); Bartlett et al. (2020) and the references therein). This was in part due to the existence of closed-form expressions for minimum ℓ_2-norm interpolators, which are particularly handy when determining the statistical risk. In contrast, the theoretical underpinnings for minimum ℓ_1-norm interpolators, despite growing interest (e.g., Ju et al. (2020); Liang and Sur (2020); Chinot et al. (2020)), remain highly inadequate and considerably more challenging to establish. Given that multiple learning algorithms are known to favor low ℓ_1-norm solutions in the over-parameterized regime (such as Rosset et al. (2004); Gunasekar et al. (2018)), understanding the statistical properties of the minimum ℓ_1-norm interpolation plays a pivotal role in unveiling the trade-offs between over-parameterization and generalization, which we seek to explore in this paper.

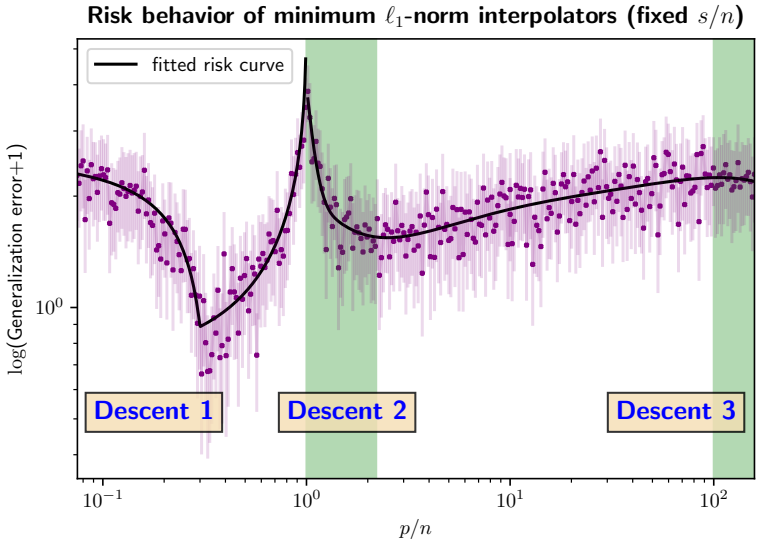

Figure 1. Triple descent in sparse linear regression (see model (1)), when the ratio of the sparsity s of the true signal and the sample size n stays fixed. More specifically, we fix $s/n = 0.3$ and $s/n \cdot M^2 = 10$ (where M is the magnitude of non-zero entries). When $s \leq p$, the true signal $\boldsymbol{\theta}^\star$ is set as an s-sparse vector. When $p < s$, we still set the true signal $\boldsymbol{\theta}^\star$ as an s-dimensional vector, while assuming we only have access to a subset of p features. We set the sample size as $n = 100$, and choose 500 values of p/n such that the $\log(p/n)$'s are uniformly spaced over $[-2, 2.2]$. In each run and for each p/n ratio, we generate a random instance and compute the minimum ℓ_1-norm interpolator and its risk. We report the average risk and error bar over these 30 independent runs for each p/n ratio. The solid line represents the fitted risk curve: when $p/n < 1$, we use the theoretical risk of least-square estimators in the current setting; when $p/n > 1$, we employ cubic spline smoothing to fit an empirical risk curve.

1.1 Motivation: a multi-descent phenomenon

An intriguing empirical phenomenon called "*double descent*" has recently emerged in the study of over-parameterized learning models (Neyshabur et al., 2014; Nakkiran et al., 2019; Belkin et al., 2018, 2019). Consider, for example, a risk curve that depicts how the generalization error varies as more parameters are added to the model. Following the classical bias-variance trade-off U-shape curve before entering the interpolation (or over-parameterized) regime (Friedman et al., 2001), the generalization error of various models descends again as one further increases the number of parameters beyond the interpolation limit. In addition, this double-descent phenomenon is also closely related to a curious observation — the non-monotonicy of risk as the model capacity grows — that has attracted much recent attention (Viering et al., 2019).

Aimed at distilling insights that help explain this phenomenon, a recent body of works studied the behavior of the minimum ℓ_2-norm interpolator in the presence of a linear model, which solidified the double-descent phenomenon for this interpolator (see, e.g., Mei and Montanari (2019); Hastie et al. (2019); Bartlett et al. (2020); Belkin et al. (2020) and the references therein). Moving beyond minimum ℓ_2-norm interpolators, empirical observations have been discussed regarding the minimum ℓ_1-norm interpolator as well; for instance, similar double descent was numerically observed in Muthukumar et al. (2020), with heuristic justification provided in Mitra (2019) based on statistical physics intuitions. Our own numerical experiments uncover even more intriguing risk behavior of the minimum ℓ_1-norm interpolator. As illustrated in Figure 1 and Figure 2, we observe "*multiple descent*" in certain parameter regimes; that is, as the model complexity continues to grow, the out-of-sample risk of the minimum ℓ_1-norm interpolator undergoes multiple phases of increase and decrease, and ultimately becomes non-increasing even as the over-parameterized ratio tends to infinity. There is, however, lack of theoretical support that elucidates this empirical observation. It remains unclear how to interpret the striking distinction in the risk behavior between the minimum ℓ_1-norm and the minimum ℓ_2-norm interpolators.

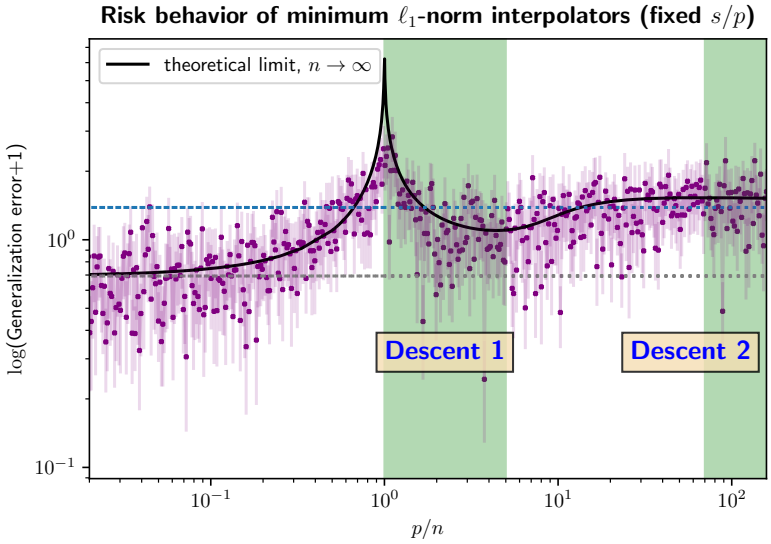

Figure 2. Multiple-descent phenomenon observed in numerical experiments. We generate data from a linear model (1) with i.i.d. Gaussian design, where parameters are set as $\sigma = 1$, SNR$:= \epsilon M^2 = 2$, and sparsity level $\epsilon = 0.01$. The sample size is fixed at $n = 100$, and choice of p's and the calculation of error bars are the same as Figure 1. The theoretical curve, predicted by results in the present paper, is shown in solid line, and the $p/n \to 0$ and $p/n \to \infty$ limits are shown in dotted lines. When $p \geq n$, two descending phases are observed here, where the first descending regime happens at the interpolation point where $n = p$, which is common for various types of models including the minimum ℓ_2-norm interpolation. The second descent appears when p/n is large enough, presenting a unique behavior for ℓ_1-norm minimization problem.

1.2 Main results and insights

In this paper, we concentrate on linear models, and investigate the generalization error (in terms of the out-of-sample squared error) of the minimum ℓ_1-norm interpolator — or equivalently, the Lasso estimator with regularization parameter approaching zero. We pursue a comprehensive understanding of such estimators in the *proportional growth* and *over-parameterized* regime, where the number of parameters p scales linearly with, but larger than, the number of samples n. Recognize that Figure 1 and Figure 2 (whose difference only lies in how the sparsity levels are chosen) exhibit very similar behavior in the over-parameterized regime. To streamline presentation and avoid repetition, we shall restrict our attention to the geometric properties of the risk curve in the setting of Figure 2. In what follows, we formulate the problem precisely, followed by a summary of our main results.

Models. Setting the stage, imagine that we have gathered n i.i.d. noisy training data drawn from a linear model

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}^\star \rangle + z_i, \qquad 1 \leq i \leq n, \tag{1}$$

where $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ is a vector composed of p unknowns, $\boldsymbol{x}_i \in \mathbb{R}^p$ stands for a (random) design vector known *a priori*, and the z_i's denote i.i.d. Gaussian noise. In addition, we consider the linear sparsity regime, where (i) the unknown signal $\boldsymbol{\theta}^\star$ is $(\epsilon \cdot p)$-sparse for some fixed constant $\epsilon > 0$, and (ii) all the $\epsilon \cdot p$ non-zero entries have magnitudes proportional to some given quantity M (to be made precise momentarily in Section 2.1).

Under the well-specified linear model, the generalization error (or out-of-sample risk) of any estimator $\widehat{\boldsymbol{\theta}}$ is defined as the expected prediction risk over a new sample data $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$:[1]

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\big[(\boldsymbol{x}_{\text{new}}^\top \widehat{\boldsymbol{\theta}} - y_{\text{new}})^2\big] \tag{2}$$

[1]The expectation is not taken w.r.t. $\boldsymbol{\theta}^\star$ here, which is however not important as the risk will converge almost surely to the expected risk in all cases considered in this paper.

where the new sample data follows the same distributions as the training data and is independent of the estimator $\widehat{\boldsymbol{\theta}}$. Our focal point is the high-dimensional asymptotics (or the large system limit), that is, we study the case when $n, p \to$ with their ratio n/p held fixed. For notational convenience, we shall often abbreviate the limiting risk as follows as long as the limit exists almost surely:

$$\mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}; \delta\big) \coloneqq \lim_{\substack{n/p=\delta \\ n,\, p\to\infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}\big). \tag{3}$$

Main findings: the risk curve of the minimum ℓ_1-norm interpolator. When it comes to the over-parametrized regime where $p > n$, the system of equations $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle$, $1 \le i \le n$ is under-determined, thus implying the existence of multiple regression parameters $\boldsymbol{\theta}$ that interpolate the training data perfectly. Among all possible interpolators, the focal point of this paper is the *minimum ℓ_1-norm intepolator*, which enjoys the smallest ℓ_1-norm as defined below

$$\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}} \coloneqq \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{subject to} \quad y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle, \quad 1 \le i \le n. \tag{4}$$

In an attempt to understand its generalization behavior, we seek to pin down the exact asymptotics of the above risk metric. Encouragingly, the large system limit of $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}})$ can be accurately pinpointed by solving a system of two nonlinear equations with two unknowns (to be formalized in Theorem 2). In turn, such risk characterizations provide a rigorous footing for the multi-descent behavior numerically observed in Figure 2, as asserted by the following theorem.

Theorem 1 (Shape of the risk curve)**.** *Suppose that $0 < \delta < 1$, and fix $n = \delta p$. Assume i.i.d. Gaussian design, i.i.d. Gaussian noise, and linear sparsity (to be made precise in Section 2.1). Then the generalization error (cf. (3)) of the minimum ℓ_1-norm interpolator (4) satisfies the following properties:*

(a) There exist two constants $1 < \eta_1 < \eta_2 < \infty$ such that $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}; \delta)$ decreases with p/n within the range $p/n \in (1, \eta_1) \cup (\eta_2, \infty)$.

(b) $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}; \delta)$ approaches the risk of the zero estimator (i.e., $\mathsf{Risk}(\mathbf{0})$) as p/n tends to infinity.

(c) For any fixed signal-to-noise ratio (to be defined precisely in (12)), there exists a constant $\epsilon^* > 0$ such that if the sparsity ratio ϵ obeys $\epsilon < \epsilon^*$, then one can find a region within the range $p/n \in (\eta_1, \eta_2)$ such that $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}; \delta)$ increases with p/n.

(d) In addition, for every given δ, there exists a threshold $\tilde{\epsilon}(\delta)$ such that $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}; \delta)$ decreases with p/n at this particular point δ as long as the sparsity ratio ϵ satisfies $\epsilon \le \tilde{\epsilon}(\delta)$.

Geometric implications and insights. Theorem 1 reveals certain geometric properties of the risk curve of the minimum ℓ_1-norm estimator in the over-parameterized regime (i.e., $p > n$). Let us take a moment to discuss the implications regarding how the risk changes in the over-parameterized ratio p/n.

- Theorem 1(a) identifies two *non-overlapping* regions within the over-parameterized regime that exhibit risk descent. Consequently, the total number of descent depends largely on the risk behavior in between these two regions.
- Theorem 1(c) indicates that the risk in between the above-mentioned two regions exhibits contrastingly different behavior depending on the sparsity ratio.
  - When the sparsity ratio is relatively small, the generalization error exhibits an intriguing "*decreasing – increasing – decreasing again*" pattern in the over-parameterized regime. This taken together with what happens in the under-parameterized regime unveils a *"triple-descent"* behavior, which matches the numerical findings in Figure 3. Interestingly, the minimum ℓ_2-norm interpolator for such a model often enjoys a double-descent behavior rather than triple descent, thus uncovering a fundamental difference between the minimum ℓ_1-norm and the minimum ℓ_2-norm interpolation.

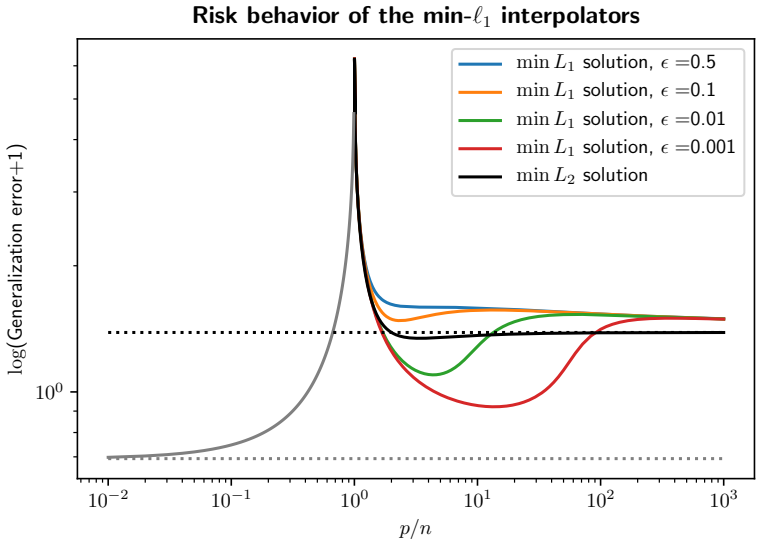

Figure 3. Theoretical risk curves for the minimum ℓ_1-norm interpolation (obtained by solving the system of equations (15)). Here, we set SNR $= 2$, and consider different values of the sparsity ratio ϵ. When $p/n < 1$, the risk curves share similar behavior as the ordinary least square estimator. When $p/n > 1$ and when ϵ drops below a certain level, the risk curves present one more descent phase. In contrast, the minimum ℓ_2-norm interpolation curve exhibits just one descent phase in the $p > n$ regime in all cases, as plotted in black.

  - In stark contrast, if the sparsity ratio is relatively large (in the sense that $\epsilon > \epsilon^\star$), the generalization error might actually decrease monotonically with p/n in the entire over-parameterized regime. If this were true, then taking it together with the classical conclusion in the under-parameterized regime would justify the double-descent behavior that has also been empirically observed in Muthukumar et al. (2020); Mitra (2019).

- In view of Theorem 1(b), the minimum ℓ_1-norm interpolator is essentially no better than a trivial estimator (i.e., the zero estimator) when the over-parameterized ratio p/n is overly large.
- Finally, Theorem 1(d) reveals that at any over-parameterized ratio, the generalization risk can be decreasing with p/n as long as the sparsity ratio is small enough.

1.3 A glimpse of our technical approach and novelty

Demonstrating the multi-descent phenomenon requires understanding the asymptotic risk of the interpolator of interest, which can be achieved by analyzing an iterative algorithm called *Approximate Message Passing* (AMP), originally proposed by Donoho et al. (2009) in the context of compressed sensing. Most relevant to our paper is the series of papers by Bayati and Montanari (2011a,b) that determined the asymptotic Lasso risk with a fixed and *strictly* positive regularization. In order to analyze the minimum ℓ_1-norm interpolator, the present work extends the AMP machinery to accommodate Lasso with the regularization parameter approaching zero, which can be accomplished by running a sequence of AMP that changes the algorithm parameters in an epoch-based manner. Noteworthily, previous analyses relied on the observation that having positive regularization encourages sparse solutions and, in turn, induces certainty restricted strong convexity around the solution. This, however, fails to capture our AMP dynamics due to the absence of positive regularization. To remedy this issue, we develop a new type of structural properties that allows one to analyze AMP iterates with changing parameters. As it turns out, the minimum ℓ_1-norm solution coincides with the fixed-point of the new AMP updates, whose risk behavior can be characterized by a new system of two nonlinear equations with two unknowns. Obtaining the exact characterization of the minimum ℓ_1-norm is beyond what prior AMP theory has to offer.

With the risk characterization in place, everything boils down to analyzing the above-mentioned nonlinear systems of equations — in particular, how its solutions vary with the aspect ratio δ. This, however, is challenging to cope with, as there is no closed-form expression of the solution points. While the prior work Miolane and Montanari (2018) studied the existence and uniqueness of the state evolution solutions, it is unclear how the solution varies with δ, particularly in the absence of strictly positive regularization. All this is addressed in the present paper via careful analysis of the first- and second-order properties of the system of equations, which constitutes much of our analysis. We expect our analysis idea to be useful to analyze other estimators such as more general M-estimators (Donoho and Montanari, 2016) or the SLOPE estimator (Su and Candes, 2016).

1.4 Other related works

Multiple descent. While the emergence of the multi-descent phenomenon in our setting is caused by the special structure of minimum ℓ_1-norm interpolators as well as the interplay between the over-parameterized ratio and the sparsity level, this phenomenon has also been observed in other settings for ℓ_2-norm minimization — albeit of different nature compared to ours. It is noteworthy that the presence of multiple descent can be caused by various other structures of the design matrix. As concrete examples, this might arise in non-isotropic linear regression where the covariance of the design matrix possesses two eigenspaces of different variance (Nakkiran et al., 2020); another possibility that leads to this phenomenon is to tweak the change points of the risk curve of the minimum ℓ_2-norm solution by carefully adding new columns (features) to the design matrix (with either standard Gaussian or Gaussian mixtures distributions) (Chen et al., 2020). Additionally, this phenomenon might also stem from the regression kernel in use. For instance, Adlam and Pennington (2020) derived the high-dimensional asymptotics for the risk curve when using neural tangent kernels in a two-layer neural network; Liang et al. (2020) studied the convergence properties of the minimum kernel-Hilbert norm interpolators under various scaling of $p = n^\alpha$, $\alpha \in (0, 1)$, and suggested possible change points (from ascent to descent) at $\alpha = \frac{1}{l+1/2}$ for every integer l. In addition, d'Ascoli et al. (2020) empirically observed the multi-descent phenomenon under the random Fourier feature model.

Minimum ℓ_1-norm solutions. In the over-parametrized regime ($p > n$), the minimum ℓ_1-norm interpolator considered herein is closely related to the problem of Basis Pursuit (BP) in the compressed sensing literature (e.g., Chen and Donoho (1994); Chen et al. (2001); Wojtaszczyk (2010); Candes and Tao (2006); Donoho (2006); Donoho et al. (2005)). In particular, the algorithm (4) has been well-established paradigm for finding a sparse solution to a noiseless linear system. When it comes to the proportional growth and linear sparsity regime in the noiseless case, Donoho and Tanner (2009); Amelunxen et al. (2014) characterized the exact phase transition boundary regarding the sample size in achieving perfect recovery. Moving to the noisy scenario, Ju et al. (2020) considered the same estimator and exhibited a double-descent phenomenon when p is exponentially larger than n. Chinot et al. (2020) studied the setting with p exceeding the order of $n \log^{1-\beta}(n)$ for some constant $\beta \in (0, 1)$, which did not focus on determining exact pre-constants and the double- or multi-descent phenomenon. Another recent work Liang and Sur (2020) studied a drastically different problem — binary classification, and pinned down exact asymptotics of the minimum ℓ_1-norm solution when the data are separable, which has intimate connection to AdaBoost.

Exact high-dimensional asymptotics. The exact asymptotic framework adopted in this work is closely related to the risk characterization of the Lasso estimator (for positive λ) that has been obtained in prior literature. In the proportional growth regime (so that p and n are comparable), the Lasso risk under i.i.d. Gaussian designs has been determined by Bayati and Montanari (2011b); Stojnic (2013); Oymak et al. (2013). In particular, the AMP machinery is a powerful tool for determining exact asymptotics in this regime, and we postpone further discussions to Section 3.1. The distributional characterization of the Lasso has been recently established by Miolane and Montanari (2018) under the i.i.d. Gaussian designs, and by Celentano et al. (2020); Bellec and Zhang (2019) under general correlated Gaussian designs, where the first two works were built upon the convex Gaussian min-max theorem. Going beyond the ℓ_1-penalty, the estimation risk of the robust regression estimators was pioneered by El Karoui (2013, 2018); Donoho and Montanari (2016) and extensively studied by, e.g., Dobriban and Wager (2018); Thrampoulidis et al. (2018); Hastie et al. (2019); Patil et al. (2021).

Approximate message passing. Inspired by statistical physics and information theory literature, AMP was first proposed as an efficient scheme to solve compressed sensing problems (Donoho et al., 2009). Bayati and Montanari (2011a); Javanmard and Montanari (2013) then rigorously proved that the dynamics of AMP can be accurately tracked by a simple small-dimensional recursive formula called the *state evolution*. This state-evolution characterization made AMP amenable as a analysis device to describe the statistical behaviors for various problems, despite that AMP is an effective algorithm on its own. The AMP algorithm and machinery has been successfully applied to a variety of problems beyond compressed sensing, including but not limited to robust M-estimators (Donoho and Montanari, 2016), SLOPE (Bu et al., 2020), low-rank matrix estimation and PCA (Montanari and Venkataramanan, 2021; Fan, 2020; Zhong et al., 2021), stochastic block models (Deshpande et al., 2015), phase retrieval (Ma et al., 2018), phase synchronization (Celentano et al., 2021), and generalized linear models (Sur et al., 2019; Sur and Candès, 2019). See Feng et al. (2021) for an accessible introduction of this machinery and its applications. Moreover, a dominant fraction of the AMP works focused on high-dimensional asymptotics (so that the problem dimension tends to infinity first before the number of iterations), except for Rush and Venkataramanan (2018) that derived finite-sample guarantees allowing the number of iterations to grow up to $O(\log n/\log\log n)$.

1.5 Notation

Here, we provide a summary of notation to be used throughout the present paper. In general, scalars are denoted by lowercase letters, vectors are represented by boldface lowercase letters, while matrices are denoted by boldface uppercase letters. For every $q \in [1,\infty]$ and any vector $\boldsymbol{x} \in \mathbb{R}^p$, we use $\|\boldsymbol{x}\|_q \coloneqq (\sum_{i=1}^p |x_i|^q)^{1/q}$ to represent the ℓ_q-norm of $\boldsymbol{x}$, and let $\|\boldsymbol{x}\|_0$ indicate the number of non-zero coordinates in $\boldsymbol{x}$. We denote by $\langle \boldsymbol{x} \rangle \coloneqq \frac{1}{p}\sum_{i=1}^p x_i$ the average of the entries of the vector $\boldsymbol{x} \in \mathbb{R}^p$. Additionally, let $\sigma_{\min}(\boldsymbol{M})$ and $\sigma_{\max}(\boldsymbol{M})$ denote respectively the minimum and the maximum singular values of a matrix $\boldsymbol{M}$.

Define $[n] \coloneqq \{1,\cdots,n\}$ for an integer n. For two functions $f(\cdot)$ and $g(\cdot)$, we often employ the convenient notation $f(\delta) \lesssim g(\delta)$ (resp. $f(\delta) \gtrsim g(\delta)$) to indicate that

$$\lim_{\delta\to\delta_0} f(\delta)/g(\delta) \le 1 \qquad (\text{resp. } \lim_{\delta\to\delta_0} f(\delta)/g(\delta) \ge 1),$$

where δ_0 is a certain limiting point that will be clear from the context. We also write $f(\delta) \sim g(\delta)$ when both $f(\delta) \lesssim g(\delta)$ and $f(\delta) \gtrsim g(\delta)$ hold true. In addition, the soft-thresholding function is defined as

$$\eta(x;\zeta) \coloneqq (|x| - \zeta)_+ \,\mathsf{sign}(x) \tag{5}$$

for any $x \in \mathbb{R}$ and a given threshold $\zeta \in \mathbb{R}^+$, where $z_+ \coloneqq \max\{z, 0\}$. Further, we let $\eta'(\cdot;\cdot)$ denote differentiation with respect to the first variable. When a function is applied to a vector, it should be understood as being applied in a component-wise manner. Following conventional notation, we denote by ∂f the sub-differential of a function f. When it comes to the ℓ_1-norm $\|\cdot\|_1$, its sub-gradient at the point $\boldsymbol{x} \in \mathbb{R}^p$ can be any vector $\boldsymbol{v} = [v_i]_{1\le i\le p}$ satisfying

$$\begin{cases} v_i = \mathsf{sign}(x_i), & \text{if } x_i \neq 0; \\ v_i \in [-1,1], & \text{if } x_i = 0. \end{cases}$$

Moreover, a function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is said to be pseudo-Lipschitz if there exists a constant $L > 0$ such that

$$|\psi(x) - \psi(y)| \le L(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2 \tag{6}$$

holds for all $x, y \in \mathbb{R}^2$. Additionally, we shall often suppress a.s. in the notation $\overset{\text{a.s.}}{=}$ for almost sure convergence if it is clear from the context.

2 Risk characterization for the minimum ℓ_1-norm interpolator

2.1 Modelling assumptions

For notational simplicity, we shall often adopt the vector and matrix notation as follows

$$\boldsymbol{z} \coloneqq [z_i]_{1\le i\le n} \in \mathbb{R}^n, \qquad \boldsymbol{X} \coloneqq [\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}, \tag{7a}$$

$$\boldsymbol{y} \coloneqq [y_i]_{1\le i\le n} = \boldsymbol{X}\boldsymbol{\theta}^\star + \boldsymbol{z} \in \mathbb{R}^n. \tag{7b}$$

To formalize the problem setting, we first impose the following assumptions on the sampling process throughout the paper.

- *Gaussian design.* We study i.i.d. Gaussian design, where each design vector is independently drawn:

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{n}\mathbf{I}_p\right), \qquad 1 \le i \le n. \tag{8}$$

  Here, the scaling factor $1/n$ is introduced merely for normalization purpose. This tractable model is widely adopted when studying the high-dimensional asymptotics of Lasso (e.g. Bayati and Montanari (2011b); Miolane and Montanari (2018); Su et al. (2017)) and has been extended to other statistical learning problems (e.g., Donoho and Montanari (2016); El Karoui (2013); Sur et al. (2019); Thrampoulidis et al. (2018)). While Gaussian design is typically not satisfied in practice, it allows for useful mathematical insights that might shed light on practical contexts.

- *Gaussian noise.* It is assumed that the noise components are independent and obey

$$\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n). \tag{9}$$

Under the above Gaussian design and Gaussian noise model, the generalization error (2) of an estimator $\widehat{\boldsymbol{\theta}}$ should be defined when $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$ is drawn from the same assumption, i.e., $y_{\text{new}} = \langle \boldsymbol{x}_{\text{new}}, \boldsymbol{\theta}^\star\rangle + z_{\text{new}}$ with $\boldsymbol{x}_{\text{new}} \sim \mathcal{N}\big(0, \frac{1}{n}\mathbf{I}_p\big)$ and $z_{\text{new}} \sim \mathcal{N}(0, \sigma^2)$. This leads to

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}) \coloneqq \mathbb{E}\big[(\boldsymbol{x}_{\text{new}}^\top\widehat{\boldsymbol{\theta}} - y_{\text{new}})^2\big] = \mathbb{E}\big[\big(\boldsymbol{x}_{\text{new}}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\big)^2\big] + \sigma^2 = \frac{1}{n}\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\big\|_2^2 + \sigma^2. \tag{10}$$

In addition, we shall make assumptions regarding how the ground truth is generated, as formalized below.

- *Linear sparsity.* Suppose that each coordinate of $\boldsymbol{\theta}^\star = [\theta_i^\star]_{1\le i\le p}$ is identically and independently drawn as follows

$$\theta_i^\star \overset{\text{i.i.d.}}{\sim} \epsilon\mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0, \tag{11}$$

  where $\mathcal{P}_c$ denotes the Dirac measure at point $c \in \mathbb{R}$, and $M > 0$ is some given quantity that determines the magnitude of a non-zero entry. In words, each coordinate is non-zero (and with magnitude $M\sqrt{\delta}$) with probability ϵ. Here, the scaling factor $\sqrt{\delta}$ is introduced solely for notational convenience, which ensures that the *signal-to-noise-ratio* (SNR) obeys

$$\text{SNR} \coloneqq \frac{\mathbb{E}\big[(\boldsymbol{x}^\top\boldsymbol{\theta}^\star)^2\big]}{\sigma^2} = \frac{\epsilon M^2}{\sigma^2}. \tag{12}$$

  When ϵ is a fixed constant, the number of non-zero coordinates concentrates around $\epsilon \cdot p$, meaning that ϵ determines the sparsity level of $\boldsymbol{\theta}^\star$. Noteworthily, a model with linear sparsity lends itself well to high-dimensional applications with only moderate degrees of sparsity (for instance, in various problems in genomics, the relevant signals are observed to be spread out across a good fraction of the genome (Boyle et al., 2017; Tam et al., 2019)).

Remark 1. It is worth noting that the linear sparsity regime often precludes consistency results in both estimation and support recovery, which is in stark contrast to the regime where the sparsity level is vanishingly small compared to the sample size (Bickel et al., 2009; Wainwright, 2009; Bühlmann and van de Geer, 2011). In fact, results featuring this regime often require an additional adjustment due to the effect of undersampling, as discussed in El Karoui et al. (2013).

2.2 Risk characterization

In order to depict the shape of the risk curve, we first characterize the precise asymptotics of $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}};\delta)$ with a fixed aspect ratio $0<\delta<1$. As alluded to previously, this is accomplished by considering sequences of instances of increasing sizes, along which the minimum ℓ_1-norm interpolator (cf. (4)) has a non-trivial limiting risk behavior.

Towards this end, we consider a more general distribution on $\boldsymbol{\theta}^\star$ by assuming that

$$\text{The empirical distribution of } \boldsymbol{\theta}^\star \text{ converges weakly to a probability measure } P_\Theta. \tag{13}$$

The following theorem determines the exact asymptotics of $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}};\delta)$ for any given $0<\delta<1$.

Theorem 2 (Risk of min ℓ_1-norm interpolation)**.** *Consider the linear model (1), and suppose that the assumptions (8), (9) and (13) hold. Consider any given $0<\delta<1$. If $\mathbb{E}[\Theta^2]<\infty$ and $\mathbb{P}(\Theta\neq 0)>0$, then the prediction risk of the minimum ℓ_1-norm interpolator obeys*

$$\lim_{\substack{n/p=\delta\\ n,p\to\infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\big) \overset{\text{a.s.}}{=} \tau^{\star 2}. \tag{14}$$

Here, $(\tau^\star,\alpha^\star)$ stands for the unique solution to the following system of equations

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[\big(\eta(\Theta+\tau Z;\alpha\tau)-\Theta\big)^2\right], \tag{15a}$$

$$\delta = \mathbb{P}\big(|\Theta+\tau Z|>\alpha\tau\big), \tag{15b}$$

where $\Theta\sim P_\Theta$, and $Z\sim\mathcal{N}(0,1)$ and is independent of Θ.

Remark 2. It is noteworthy that Theorem 2 is completely general regarding the distribution of $\boldsymbol{\theta}^\star$ as long as its empirical distribution converges to a fixed measure; in particular, it does not require Θ to follow the sparse distribution specified in the expression (11).

First, there exists a unique solution pair to the set of equations (15) as asserted by Proposition B.1 (see Section B for more details). Experienced readers who are familiar with literature on Lasso shall immediately recognize the similarity between these equations and the ones used to determine the Lasso risk in the proportional regime (Bayati and Montanari, 2011b). We shall elaborate a bit more on their connections and differences in Section 2.3.

We now pause to interpret the above result. The risk of the minimum ℓ_1-norm interpolator — when the ratio n/p is held fixed — converges to a quantity $\tau^{\star 2}$, which is a function of (σ,δ,P_Θ) and can be determined by solving a system of two nonlinear equations with two unknowns. At a high level, the equation (15a) indicates that $\tau^\star>\sigma$, which can be viewed as variance inflation as a result of undersampling. In addition, $\tau^\star$ taken together with the other parameter $\alpha^\star$ controls the sparsity level of $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$. In fact, as can be seen from the equation (15b) and our analysis, we have

$$\lim_{\substack{n/p=\delta\\ n,p\to\infty}} \frac{1}{p}\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_0 \overset{\text{a.s.}}{=} \mathbb{P}\big(|\Theta+\tau^\star Z|>\alpha^\star\tau^\star\big)=\delta,$$

which is essentially saying that the support size of $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ converges to n in the limit. The proof of Theorem 2 is established via analyzing a sequence of *Approximate Message Passing* (AMP) updates with careful choices of parameters, such that the minimum ℓ_1-norm solution is the fixed point of these updates. The state evolution formula that characterizes the large n limit for each iterate is derived, and its large t limit corresponds to the risk of the minimum ℓ_1-norm solution. The readers are referred to Section 3.1 for details.

Multi-descent phenomenon. Having obtained an exact characterization of $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}})$ in a general manner, we are ready to specialize Theorem 2 to the ground-truth distribution (11) and examine how $\tau^\star$ changes as a function of the aspect ratio δ. Specifically, if denote $\nu\coloneqq M/\tau$, then the equations (15) simplify to

$$1=\frac{\nu^2}{M^2}\sigma^2+\frac{\epsilon}{\delta}\mathbb{E}\left[\big(\eta(\sqrt{\delta}\nu+Z;\alpha)-\sqrt{\delta}\nu\big)^2\right]+\frac{1-\epsilon}{\delta}\mathbb{E}\left[\eta^2(Z;\alpha)\right] \tag{16a}$$

$$\delta = \epsilon \mathbb{P}\big(|\nu\sqrt{\delta} + Z| > \alpha\big) + (1-\epsilon)\mathbb{P}(|Z| > \alpha) \tag{16b}$$

in the presence of the distribution (11). From equation set (16), we can readily examine how $\tau^\star$ varies with δ, which is the content of Section 3.2 (along with the corresponding appendix). In particular, we can use (16) to demonstrate that: the risk curve undergoes a phase transition in terms of the sparsity level ϵ — as summarized in Theorem 1 — such that the curve transitions from a single descent to multiple descent in the over-parameterized regime. To the best of our knowledge, this provides the first theoretical justification for the multiple-descent phenomenon associated with the minimum ℓ_1-norm interpolator, and might shed light on understanding the behavior of other interpolators such as the M-estimators with a general family of objective functions.

Comparisons with ridgeless regression. Hastie et al. (2019) investigated the risk behavior of the ridge estimator when the penalized parameter λ tends to zero — which corresponds to the minimum ℓ_2-norm interpolator in the over-parameterized regime — and solidified a double-descent phenomenon as one increases the over-parameterized ratio p/n. To facilitate comparisons to their results, we first translate the results in Hastie et al. (2019) using our notation. Specifically, the generalization error of the minimum ℓ_2 interpolator — denoted by $\widehat{\boldsymbol{\theta}}^{\mathsf{Int},\ell_2}$ — obeys

$$\lim_{\substack{n/p=\delta \\ n,p\to\infty}} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int},\ell_2}) \overset{\text{a.s.}}{=} \begin{cases} \frac{\delta}{\delta-1}\sigma^2, & \text{if } \delta > 1 \\ \epsilon M^2(1-\delta) + \frac{1}{1-\delta}\sigma^2, & \text{if } \delta < 1 \end{cases} \tag{17}$$

under the model (11). By calculating the derivative of the right-hand side of (17) w.r.t. δ, one can easily demonstrate that the exact asymptotics of $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int},\ell_2})$ decays with[2] $1/\delta$ when $\epsilon \le \sigma^2/M^2$; otherwise, if $\epsilon > \sigma^2/M^2$, then the risk curve undergoes a decreasing phase before hitting the point associated with $\delta = 1 - \frac{\sigma}{\sqrt{\epsilon}M}$, and starts to increase with $1/\delta$ afterward. Next, we single out a few key differences between their results and ours in Theorem 1.

- The current paper considers the case where the sparsity ratio of $\boldsymbol{\theta}^\star$ is held fixed across different random instances of increasing dimension, with the SNR frozen to be $\epsilon M^2/\sigma^2$. The role of over-parametrization is studied when the minimum ℓ_1-norm estimator (which naturally promotes sparse solutions) is fitted with full model dimension p.

  In contrast, Hastie et al. (2019) studied the case where the underlying signal $\boldsymbol{\theta}^\star$ has a bounded ℓ_2-norm and potentially dense.

- Interestingly, Theorem 1 suggests that the minimum ℓ_1-norm interpolator often exhibits more than two descent, thus revealing a fundamental difference between these two types of interpolation.

- There exists a convenient closed-form expression for the minimum ℓ_2-norm interpolator, which assists in characterizing the precise asymptotics (i.e., one can decompose the risk formula into bias and variance terms, and pin down each term with the aid of random matrix theory). Unfortunately, the minimum ℓ_1-norm interpolator does not admit a concise closed-form expression, thus making it considerably more challenging to analyze. In light of this, Section 3.2 is devoted to the analysis of the above-mentioned nonlinear system of equations, with the aim of determining (local) monotoncity of the corresponding quantities of interest.

2.3 Connections to the Lasso estimator

Apparently, the minimum ℓ_1-norm interpolator (4) is closely related to the classical Lasso estimator studied extensively in high-dimensional statistics (Tibshirani, 1996). Given a positive regularization parameter $\lambda > 0$, the Lasso estimates the regression coefficients by solving the following optimization problem

$$\widehat{\boldsymbol{\theta}}_\lambda := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}. \tag{18}$$

[2]Following the convention, we study the relation regarding $1/\delta = p/n$ instead of $\delta = n/p$.

As a consequence, the minimum ℓ_1-norm interpolator corresponds to the limit of $\widehat{\boldsymbol{\theta}}_\lambda$ when taking λ to zero.

Several prior works have attempted to characterize the exact asymptotics of the Lasso risk $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda)$ in the proportional regime. Specifically, it has been proven that for any given $\lambda > 0$, $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda)$ converges to a non-trivial limit $\tau^\star(\lambda)$. Here, $(\tau^\star(\lambda), \alpha^*(\lambda))$ represents the solution pair to the following set of nonlinear equations

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[\big(\eta(\Theta + \tau Z, \alpha\tau) - \Theta\big)^2\right], \tag{19a}$$

$$\lambda = \alpha\tau\left(1 - \frac{1}{\delta}\mathbb{P}\big(|\Theta + \tau Z| > \alpha\tau\big)\right), \tag{19b}$$

where $\Theta \sim P_\Theta$ and $Z \sim \mathcal{N}(0,1)$ are independent random variables. The interested reader can consult Bayati and Montanari (2011b); Miolane and Montanari (2018); Celentano et al. (2020), which determined the Lasso risk using either the AMP machinery or the convex Gaussian min-max theorem.

As can be easily seen, the system of equations (19) bears much resemblance to (15). More precisely, by directly setting λ to 0, the equation set (19) reduces to the one in (15). In other words, one has

$$\lim_{\lambda\to 0}\lim_{\substack{n/p=\delta\\ n,p\to\infty}} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda) \overset{\text{a.s.}}{=} \lim_{\lambda\to 0}[\tau^\star(\lambda)]^2 = \tau^{\star 2}. \tag{20}$$

where $\tau^\star$ denotes the quantity in Theorem 2, and the last identity holds under certin continuity assumptions w.r.t. the equations (19).

Intuitively, Theorem 2 can be directly established if it is legitimate to switch the order of limits between λ and p on the left hand side of expression (20), given that the minimum ℓ_1-norm interpolator is the limit of the Lasso by taking λ to zero. However, formally establishing the validity of exchanging limits is quite challenging, since doing so normally requires the loss function being strongly convex (at least locally strongly convex around the solution point). Such a strong convexity property, however, is lacking in our problem structure when λ is taken to zero. In fact, this presents a major roadblock to directly applying the established AMP theory for the Lasso estimator.

Fortunately, we can directly argue that exchanging the two limits leads to the same result, as formalized in the proposition below. The proof of this result is postponed to Section A.

Proposition 1 (The Lasso limit when $\lambda \to 0$)**.** *In the setting of Theorem 2, the Lasso risk obeys the following asymptotically exact characterization:*

1. *When $\delta < 1$, the asymptotic Lasso risk converges to the risk of min ℓ_1-norm interpolator (4):*

$$\lim_{\lambda\to 0}\lim_{\substack{n/p=\delta\\ n,p\to\infty}} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda) \overset{\text{a.s.}}{=} \tau^{\star 2},$$

   with $\tau^\star$ being the solution to the system of equations (15).

2. *When $\delta > 1$, the asymptotic Lasso risk converges to the risk of the ordinary least-square solution, namely,*

$$\lim_{\lambda\to 0}\lim_{\substack{n/p=\delta\\ n,p\to\infty}} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda) \overset{\text{a.s.}}{=} \frac{\delta}{\delta - 1}\sigma^2.$$

In words, the above result reveals that: while the connection between the set of equations (19) and the Lasso risk was previously only shown for a positive λ, such exact asymptotics continue be valid even in the limit when λ approaches zero.

3 Key analysis

This section presents the key ideas for proving our main results. We start by presenting the proof strategy for Theorem 2, which is built upon the recently developed approximate message passing machinery. It is then followed by the proof of Theorem 1 that characterizes the geometric properties of the risk curve.

3.1 Key analysis tool: approximate message passing

The major technical enabler for proving Theorem 2 lies in the recent development of an iterative algorithm called the *Approximate Message Passing* (AMP) algorithms. As mentioned previously, Bayati and Montanari (2011b) employed AMP to pin down the risk of the Lasso estimator with positive regularization. Motivated by this line of works, this paper resorts to the AMP technique as a proof device towards understanding the risk behavior of the minimum ℓ_1-norm interpolators.

For our purpose, we need to generalize the original AMP updates (Bayati and Montanari (2011b)) — which were designed to solve a single Lasso problem in the large-system limit — to approximate a sequence of Lasso problems with changing (and converging) regularization parameters. To better illustrate this idea, we shall first provide a brief review of how AMP is invoked to solve a single Lasso problem, followed by a generalization of this framework to accommodate the minimum ℓ_1-norm interpolator.

3.1.1 AMP for the Lasso estimator

AMP updates for Lasso. Recall that the soft-thresholding function is defined in expression (5) and $\langle\cdot\rangle$ denotes the average of the coordinates for the target vector. When initialized at $\boldsymbol{\theta}^0 = \mathbf{0}$ and $\boldsymbol{z}^{-1} = \mathbf{0}$, the AMP algorithm proceeds recursively in the following fashion

$$\boldsymbol{\theta}^{t+1} = \eta(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\theta}^t; \zeta_t); \tag{21a}$$

$$\boldsymbol{z}^t = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^t + \frac{1}{\delta}\boldsymbol{z}^{t-1}\left\langle \eta'_{t-1}(\boldsymbol{X}^\top \boldsymbol{z}^{t-1} + \boldsymbol{\theta}^{t-1}; \zeta_{t-1})\right\rangle. \tag{21b}$$

Here, $\{\zeta_t\}_{t=0}^{\infty}$ is an appropriate sequence of scalars to be selected. To approximate the Lasso solution with positive $\lambda > 0$ (defined in (18)), Bayati and Montanari (2011b) showed that it suffices to set

$$\zeta_t = \alpha^\star(\lambda)\cdot\tau_t(\lambda), \quad \text{for all } t \geq 0, \tag{22}$$

where $\alpha^\star(\lambda)$ is taken as the corresponding solution to the fixed-point equation (19) and $\tau_t(\lambda)$ shall be specified momentarily. Given this choice of parameters, Bayati and Montanari (2011b, Theorem 1.8) proved that the corresponding AMP update $\boldsymbol{\theta}^t$ converges to the Lasso solution in the following sense: as long as $\mathbb{E}[\Theta^2] < \infty$ and $\mathbb{P}(\Theta \neq 0) > 0$, it holds that

$$\lim_{t\to\infty}\lim_{\substack{n/p=\delta\\ n,p\to\infty}} \frac{1}{p}\|\boldsymbol{\theta}^t - \widehat{\boldsymbol{\theta}}_\lambda\|_2^2 \overset{\text{a.s.}}{=} 0. \tag{23}$$

We emphasize that this convergence result requires taking the limit of the model dimensions before taking the limit of the iteration steps; hence, it should be understood as high-dimensional asymptotics. Equipped with this result, one is able to study the limiting performance of Lasso via the AMP iterations at each fixed step t; the latter is made possible by the state evolution characterization to be introduced below.

State evolution. Consider the AMP procedure (21) with an arbitrary sequence of thresholds $\{\zeta_t\} > 0$. The state evolution sequence $\{\tau_t^2\}_{t=0}^{\infty}$ is a one-dimensional iteration sequence, recursively defined for all $t \geq 0$ as follows

$$\tau_{t+1}^2 = \mathsf{F}(\tau_t^2, \zeta_t) \tag{24a}$$

$$\text{where}\quad \mathsf{F}(\tau^2, \zeta) := \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[\left[\eta(\Theta + \tau Z; \zeta) - \Theta\right]^2\right] \tag{24b}$$

with initialization $\tau_0^2 = \sigma^2 + \mathbb{E}[\Theta^2]/\delta$. Here, Θ is the distribution of the true signal, and $Z \sim \mathcal{N}(0,1)$ is independent of Θ. The above sequence is known to characterize the limiting variance of the AMP recursion, as formalized by the following result.

Proposition 2 (Theorem 1.1, Bayati and Montanari (2011b))**.** *Consider the linear model (1) and i.i.d. Gaussian design. If $\mathbb{E}[\Theta^2] < \infty$ and $\mathbb{P}(\Theta \neq 0) > 0$, then for any positive sequence $\{\zeta_t\}$ and any pseudo-Lipschitz function ψ, it holds that*

$$\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\psi(\theta_i^{t+1}, \theta_i^\star) \overset{\text{a.s.}}{=} \mathbb{E}\big[\psi\big(\eta(\Theta + \tau_t Z; \zeta_t), \Theta\big)\big],$$

where $Z \sim \mathcal{N}(0,1)$ *is independent of* Θ.

In words, this proposition asserts that the coordinates of $\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\theta}^t$ have roughly the same distribution as $\Theta + \tau_t Z$. Taking $\psi(x,y) = (x-y)^2$ and combining this with the expression (24b) indicate that: the asymptotic risk of the AMP in the t-th iteration is characterized by τ_t^2. Indeed, the state evolution τ_t^2 quantifies how this asymptotic risk evolves with the iteration count. If we let the iteration number t tend to infinity, then τ_t^2 converges to a nonzero limit — i.e., the solution to the system of equations (19) — which is precisely the limiting risk for the Lasso estimator by virtue of the property (23).

3.1.2 AMP for the minimum ℓ_1-norm interpolator

As discussed above, when AMP adopts the choice of $\zeta_t = \alpha^\star(\lambda) \cdot \tau_t(\lambda)$, then in each iteration, it makes progress towards the Lasso solution in the presence of a positive regularization parameter λ. Intuitively, one can run AMP in an epoch-based manner, and gradually reduce the value of λ by taking a vanishing sequence of $\{\lambda_t\}$ and set $\zeta_t = \alpha^\star(\lambda_t) \cdot \tau_t(\lambda_t)$. Heuristically, each epoch solves a Lasso problem approximately with parameter λ_t and, in the end, one can recover the minimum ℓ_1-norm interpolator in the limit. Similar heuristics have been pointed out by Donoho et al. (2010) without a rigorous argument.

It turns out this intuition can be solidified as long as one selects the sequence $\{\lambda_t\}$ appropriately. Let us now describe our choice of the $\{\lambda_t\}$ sequence, and use them to construct $\{\zeta_t\}$ in the AMP updates.

Choice of $\{\zeta_t\}$ in our setting. Our first step is to construct a positive sequence of $\{\lambda_t\}$ satisfying the following assumption:

Assumption 1. For every $t = 1, 2, \ldots$, define $\Lambda_t \coloneqq \sum_{s=1}^t \lambda_s$. We assume that $\{\lambda_t\}_{t=1}^\infty$ satisfies the following conditions:

- $\lim_{t\to\infty} \lambda_t = 0$ and $\lim_{t\to\infty} \lambda_t/\lambda_{t+1} = 1$;
- $\sum_{j=t/2}^t \lambda_j \geq c \log t$ for every constant c and sufficiently large t;
- The following two sequences are summable for every constant c,

$$\sum_{t=1}^{+\infty} \exp\{-c\Lambda_t\} < \infty, \qquad \text{and} \qquad \sum_{t=1}^{+\infty} \sqrt{l_t} < \infty, \tag{25}$$
$$\text{where} \quad l_t \coloneqq \sum_{s=1}^t \big|\lambda_s - \lambda_{s+1}\big| \exp\big(-c[\Lambda_t - \Lambda_s]\big).$$

In words, Assumption 1 requires that $\{\lambda_t\}$ converges to 0, but the convergence rate should be slow enough. We shall provide an example of $\{\lambda_t\}$ satisfying this assumption in Section E.1. As will be made clear from the proof, having $\lambda_t \to 0$ guarantees that the solution to the system of equations converges to the minimum ℓ_1-norm solution, whereas other conditions ensure that the AMP iterates do not experience drastic changes in adjacent iterations. In particular, the difference between the support sets of consecutive iterates can be properly controlled.

With this choice of $\{\lambda_t\}$ sequence, we define a series of nonlinear systems of equations with two unknowns, indexed by t as follows:

$$\begin{aligned} \tau^2 &= \mathsf{F}(\tau^2, \alpha\tau), \\ \lambda_t &= \alpha\tau \left(1 - \frac{1}{\delta}\mathbb{E}\big[\eta'(\Theta + \tau Z; \alpha\tau)\big]\right), \end{aligned} \tag{26}$$

where the function $\mathsf{F}(\cdot,\cdot)$ is specified in expression (24b). As usual, Z is a standard Gaussian random variable that is independent of Θ. Recognizing the existence and uniqueness property shown in Section B, we can guarantee that the equation set (26) yields a unique solution pair, which shall be denoted by $(\alpha_t^\star, \tau_t^\star)$. Further, we define the threshold ζ_t for our AMP updates (21) as follows

$$\zeta_t \coloneqq \alpha_t^\star \cdot \tau_t \qquad \text{for all } t > 0, \tag{27}$$

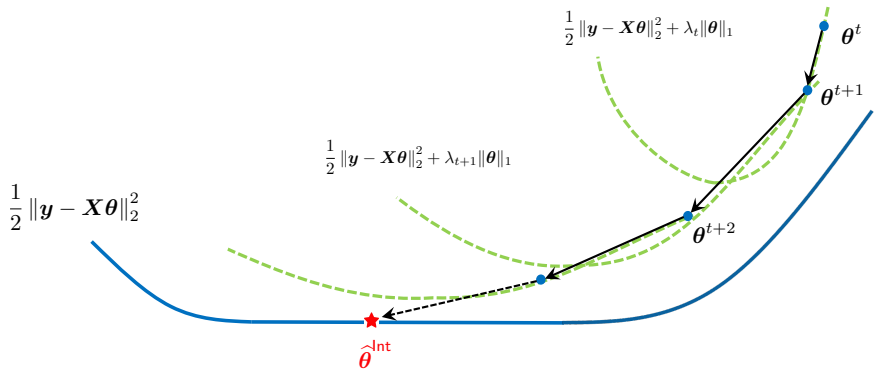

Figure 4. Illustration of the AMP updates for the minimum ℓ_1-norm interpolator. At each step, $\boldsymbol{\theta}^{t+1}$ is computed as in the expression (21a), with ζ_t chosen according to (27). The plateau of the blue curve stands for all the interpolators that satisfy $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}$, among which $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ has the smallest ℓ_1-norm. The curves in green stand for the Lasso loss functions (with parameter λ_t changing with t) where AMP aims to move towards its minimizer in each t.

where $\zeta_0 \coloneqq 1$ and τ_t corresponds to the state evolution formula provided in the expression (24a). In view of the correspondence between the AMP updates and the Lasso estimator, iteration t of our AMP updates takes a step towards approximating the Lasso estimator with parameter λ_t. As λ_t converges to zero, the iteration procedure has the minimum ℓ_1-norm interpolator as a limiting point. The informal intuition is illustrated in Figure 4.

We are now ready to state our main result on the risk of the min ℓ_1-norm solution.

Theorem 3. *Consider the linear model (1) and i.i.d. Gaussian design. If $\mathbb{E}[\Theta^2] < \infty$ and $\mathbb{P}(\Theta \neq 0) > 0$, then for any pseudo-Lipschitz function ψ, one has*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \psi(\widehat{\theta}_i^{\mathsf{lnt}}, \theta_i^\star) \overset{\text{a.s.}}{=} \mathbb{E}\big[\psi(\eta(\Theta + \tau^\star Z; \alpha^\star\tau^\star), \Theta)\big], \tag{28}$$

where $Z \sim \mathcal{N}(0,1)$ is independent of Θ. Here, $(\alpha^\star, \tau^\star)$ is the solution to the following equations

$$\tau^2 = \mathsf{F}(\tau^2, \alpha\tau); \quad \frac{1}{\delta}\mathbb{E}\big[\eta'(\Theta + \tau Z; \alpha\tau)\big] = 1. \tag{29}$$

We now point out an immediate consequence of Theorem 3. In view of the pseudo-Lipschitz property of the function $\psi(a,b) = (a-b)^2$, we can obtain Theorem 2 as a corollary, namely, the limiting risk of the minimum ℓ_1-norm solution obeys

$$\lim_{p\to\infty} \frac{1}{p}\big\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}} - \boldsymbol{\theta}^\star\big\|_2^2 \overset{\text{a.s.}}{=} \tau^{\star 2}. \tag{30}$$

This enables us to study the risk curve by examining the equations (29).

Proof ideas. Before proceeding, let us highlight several key challenges and differences in this part of the proof in comparison to Bayati and Montanari (2011b). The complete details are deferred to Section C. As already mentioned, we look at a sequence of AMP updates, each targeting at solving a Lasso problem with a different regularization parameter λ_t that obeys $\lim_{t\to\infty}\lambda_t = 0$. In the fixed λ scenario, it is known that even if $p > n$, the loss function around the Lasso estimate enjoys certain restricted strongly convexity. This is, however, not the case for the minimum ℓ_1-norm interpolator, whose support size equals n; this implies that the condition number of $\boldsymbol{X}^\top\boldsymbol{X}$ (restricted to the support) might be very large. Consequently, it calls for the development of a new structural property tailored to the minimum ℓ_1-norm solution, as we shall detail in Lemma C.2.

Moreover, for each λ_t, the AMP iterations are contractive towards different fixed points (i.e., minimizers of different Lasso problems). One thus needs to investigate how the pesudo-state evolution point $\tau_t^\star$ varies with

the iteration number t. In addition, to demonstrate that AMP converges to the new system of equations as specified in (15), at a high level, we construct some distance measure between $\boldsymbol{\theta}_{t+1}$ and $\boldsymbol{\theta}_t$ so as to guarantee that

$$\mathsf{dist}(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) \le \exp(-\lambda_t) \cdot \mathsf{dist}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) + c|\lambda_t - \lambda_{t+1}|. \tag{31}$$

In the case of a fixed λ, the above relation simplifies to $\mathsf{dist}(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) \le \exp(-\lambda) \cdot \mathsf{dist}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1})$, which means that $\mathsf{dist}(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t)$ converges linearly and $\boldsymbol{\theta}_t$ converges to the corresponding limit. In contrast, the second term on the right-hand side of (31) reflects the price one needs to pay when λ_t varies across iterations. Assumption 1 is imposed to help ensure that these errors — albeit accumulated over time — stay bounded.

3.2 Analysis ingredients for the risk curve

Thus far, we have demonstrated that the risk curve of the minimum ℓ_1-norm interpolator can be characterized by the solutions to the system of equations (15). In order to analyze the shape of the risk curve and establish Theorem 1, this section takes a close look at the geometric properties of these solutions. For ease of exposition, let us assume without loss of generality that $\sigma^2 = 1$ throughout the proof; clearly, having a different value of σ^2 does not change the shape of the curve as long as the SNR remains unchanged.

Roadmap of the proof. We start by providing a roadmap of our proof. To begin with, it is challenging to track the behavior of the solution $\tau^\star$ directly in the original form of the equations (15); in fact, $\tau^\star$ blows up as $\delta \to 1$. Hence, we find it more convenient to work with a new parameter $\nu := M/\tau$ (resp. $\nu^\star := M/\tau^\star$) that leads to alternative versions of Theorem 1 and (15). With this change of variables in place, it suffices to study how $\nu^\star$ behaves as one varies δ. Towards this end, we first eliminate the parameter α and express ν purely as a function of δ. We then proceed to analyze the derivative of $\nu^\star(\delta)$ using the implicit function theorem, with a special focus on the sign of $\nu^{\star\prime}(\delta)$ when δ is close to 0 or 1, as well as when $\epsilon \to 0$. As we shall argue momentarily, these steps suffice in establishing Theorem 1.

An equivalent formulation. As mentioned above, let us denote $\nu := M/\tau$, and define two functions of (ν, δ, α) as follows

$$F_1(\nu, \delta, \alpha) := \epsilon \mathbb{P}\left(\left|\nu\sqrt{\delta} + Z\right| > \alpha\right) + (1-\epsilon)\mathbb{P}\left(|Z| > \alpha\right) - \delta, \tag{32a}$$

$$F_2(\nu, \delta, \alpha) := \frac{\nu^2}{M^2} - 1 + \frac{\epsilon}{\delta}\mathbb{E}\left[\left(\eta(\sqrt{\delta}\nu + Z; \alpha) - \sqrt{\delta}\nu\right)^2\right] + \frac{1-\epsilon}{\delta}\mathbb{E}\left[\eta^2(Z; \alpha)\right], \tag{32b}$$

where $Z \sim \mathcal{N}(0,1)$ and is independent of $\Theta \sim P_\Theta$. Under our assumptions on Θ (cf. (11)), solving the equations (15a) and (15b) can be accomplished by first finding the solutions $\big(\nu^\star(\delta), \alpha^\star(\delta)\big)$ to

$$\begin{cases} F_1(\nu, \delta, \alpha) = 0, \\ F_2(\nu, \delta, \alpha) = 0, \end{cases} \tag{33}$$

and then mapping $\nu^\star$ back to $\tau^\star$.

3.2.1 Step 1: existence of the mapping $\nu^\star(\delta)$

We now attempt to eliminate the variable α in (33), and expressing $\nu^\star$ as a function of δ. For any $\delta \in (0,1)$ and $\nu > 0$, direct computation of the derivative of the function $F_1(\nu, \delta, \alpha)$ yields

$$\nabla_\alpha F_1(\nu, \delta, \alpha) = -\epsilon\left[\phi(\alpha - \sqrt{\delta}\nu) + \phi(\alpha + \sqrt{\delta}\nu)\right] - 2(1-\epsilon)\phi(\alpha) < 0.$$

It is also straightforward to calculate the limiting values

$$\lim_{\alpha \to 0^+} F_1(\nu, \delta, \alpha) = 1 - \delta > 0; \qquad \lim_{\alpha \to +\infty} F_1(\nu, \delta, \alpha) = -\delta < 0.$$

Based on the above observations, given any δ and ν, the function $F_1(\nu, \delta, \alpha)$ is monotonically non-increasing in α, and can take both positive and negative values within the interval $(0, \infty)$. As a result, there exists a mapping from $(\nu, \delta) \to \alpha$ that satisfies $F_1(\nu, \delta, \alpha) = 0$; with an abuse of notation, we often denote this function as $\alpha(\nu, \delta)$. Substitution into the function F_2 allows us to define

$$F_3(\nu, \delta) \coloneqq F_2(\nu, \delta, \alpha(\nu, \delta)). \tag{34}$$

Here, the function F_3 depends solely on the two parameters (ν, δ). Armed with the derivations above, solving (33) comes down to finding a solution to $F_3(\nu, \delta) = 0$.

By construction of the function F_3, we know that the solutions to $F_3(\nu, \delta) = 0$ correspond to the solutions to the system of equations (15). As we shall demonstrate in Proposition B.1, for every $0 < \delta < 1$, there exists a unique pair of (τ, α) satisfying (15); therefore, $F_3(\nu, \delta) = 0$ yields a unique solution for every $0 < \delta < 1$ — which shall be denoted by $\nu^\star(\delta)$ in the sequel. Correspondingly, the solution pair for (15) shall be written as $(\tau^\star(\delta), \alpha^\star(\delta))$ where $\alpha^\star(\delta) \coloneqq \alpha(\tau^\star(\delta), \delta)$. We also note that since both F_1 and F_2 are smooth functions with bounded derivatives w.r.t. all parameters, $\nu^{\star\prime}(\delta)$ exists and is continuous.

3.2.2 Step 2: derivative of $\nu^\star(\delta)$

With the mapping $\nu^\star(\delta)$ in place, we can translate Theorem 1 into statements about $\nu^{\star\prime}(\delta)$. Before doing so, recall that the risk incurred by using $\boldsymbol{\theta} = \mathbf{0}$ as the estimator satisfies

$$\mathsf{Risk}(\mathbf{0}) = \mathbb{E}\left[(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^\star \rangle + z_i)^2\right] = 1 + \frac{1}{n}\|\boldsymbol{\theta}^\star\|_2^2 \xrightarrow{\text{a.s.}} 1 + \epsilon M^2 =: \tau_0^2. \tag{35}$$

Let us define the corresponding value of ν as $\nu_0 \coloneqq M/\tau_0$. Formally, to prove the first two claims in Theorem 1, it suffices to establish the following proposition.

Proposition 3. *In the setting of Theorem 1, for $\delta \in (0, 1)$, $\nu^\star(\delta)$ satisfies the following properties:*

1. $\lim_{\delta \to 0^+} \nu^\star(\delta) = \nu_0$;
2. *There exist two constants $0 < \delta_1, \delta_2 < 1$ such that when $0 < \delta < \delta_1$ and $\delta_2 < \delta < 1$, $\nu^{\star\prime}(\delta) < 0$.*

Clearly, if Proposition 3 were valid, then the first two claims in Theorem 1 would follow immediately by invoking the change of variables $\tau = M/\nu$. Now we discuss how to establish this proposition. For notational simplicity, we use $\nu^\star$ and $\alpha^\star$ to denote the unique solution to (15a) and (15b) for any $\delta \in (0, 1)$, which should be understood as $\nu^\star(\delta)$ and $\alpha^\star(\delta)$. To begin with, recognizing the fact that $F_1(\nu^\star, \delta, \alpha^\star) = 0$, the implicit function theorem implies that

$$\nabla_\nu \alpha(\nu^\star, \delta) = -\frac{\nabla_\nu F_1(\nu, \delta, \alpha)}{\nabla_\alpha F_1(\nu, \delta, \alpha)}\bigg|_{(\nu^\star, \delta, \alpha^\star)}; \quad \nabla_\delta \alpha(\nu^\star, \delta) = -\frac{\nabla_\delta F_1(\nu, \delta, \alpha)}{\nabla_\alpha F_1(\nu, \delta, \alpha)}\bigg|_{(\nu^\star, \delta, \alpha^\star)}. \tag{36}$$

We are now ready to derive an explicit expression of $\nu^{\star\prime}(\delta)$. A little algebra leads to

$$\begin{aligned}\nu^{\star\prime}(\delta) = -\frac{\nabla_\delta F_3(\nu, \delta)}{\nabla_\nu F_3(\nu, \delta)}\bigg|_{(\nu^\star, \delta)} &= -\frac{\nabla_\delta F_2(\nu, \delta, \alpha) + \nabla_\delta \alpha(\nu, \delta) \nabla_\alpha F_2(\nu, \delta, \alpha)}{\nabla_\nu F_2(\nu, \delta, \alpha) + \nabla_\nu \alpha(\nu, \delta) \nabla_\alpha F_2(\nu, \delta, \alpha)}\bigg|_{(\nu^\star, \delta, \alpha^\star)} \\ &= -\frac{\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1}{\nabla_\nu F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\nu F_1}\bigg|_{(\nu^\star, \delta, \alpha^\star)},\end{aligned} \tag{37}$$

where the second equality invokes the relation (36). This expression plays a crucial role in our subsequent analysis in understanding how $\tau^\star$ changes with δ.

In order to establish Proposition 3, we gather in the following two lemmas some key facts on the limiting behavior of $\nu^{\star\prime}(\delta)$ and $\nu^\star(\delta)$, when $\delta \to 0^+$ and $\delta \to 1^-$, respectively. All of these are stated with the assumptions of Theorem 1 imposed, with the proofs deferred to Section D.1.

Lemma 1. *When $\delta \to 0^+$, the numerator and denominator in (37) satisfy respectively the following properties*

$$\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1 \sim -2\alpha^{\star -3}, \tag{38a}$$

$$\nabla_\nu F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\nu F_1 \sim -4\nu_0^{-1}\phi(\alpha^\star), \tag{38b}$$

where all the partial derivatives of F_1 and F_2 are evaluated at the point $(\nu^\star, \delta, \alpha^\star)$. Additionally, it holds that

$$\lim_{\delta \to 0^+} \nu^\star = \nu_0. \tag{39}$$

Lemma 2. *When $\delta \to 1^-$, it satisfies that*

$$\lim_{\delta \to 1^-} \nu^{\star\prime}(\delta) = -\infty. \tag{40}$$

Note that the property (39) in Lemma 1 validates the first claim in Proposition 3 directly. Therefore, to prove Proposition 3, we are only left with verifying the second claim in Proposition 3. In view of Lemma 1, it is guaranteed that as $\delta \to 0^+$, both the denominator and the numerator in (37) are negative. By continuity of $\nu^{\star\prime}(\delta)$, there exists some $\delta_1 > 0$ such that when $0 < \delta < \delta_1$, one has $\nu^{\star\prime}(\delta) < 0$. Finally, Lemma 2 immediately suggests that one can find $\delta_2 > 0$ such that: when $\delta_2 < \delta < 1$, one has $\nu^{\star\prime}(\delta) < 0$. Taking these properties collectively concludes the proof of Proposition 3.

3.2.3 Step 3: limit behavior for the case with $\epsilon \to 0$

Finally, let us move on to establishing the third and fourth claims of Theorem 1. Specifically, fixing some $\delta > 0$, we shall study how the risk limit $\tau^\star$ behaves as ϵ varies, particularly as it tends to zero. Thus far, we have focused on the case when both ϵ and M are regarded as fixed constants while the value of δ varies; in this case, the analyses were primarily performed w.r.t. $\nu^{\star\prime}(\delta)$, since studying $\nu^\star(\delta)$ and $\nu^\star(\delta)/M$ are equivalent when M is taken to be a fixed constant. However, in the case when we fix SNR (namely, ϵM^2) as opposed to M, one has $M \to \infty$ as $\epsilon \to 0$, and hence studying $\nu^\star(\delta)$ and studying $\nu^\star(\delta)/M$ are no longer equivalent. As a result, we need to analyze $\nu^\star(\delta)/M$ directly, that is, to examine the behavior of $\nu^\star(\delta)/M$ and $\nu^{\star\prime}(\delta)/M$ in a fixed-SNR regime as ϵ approaches zero.

The readers shall also bear in mind that we always focus on the derivative of the risk of the model (11) with given (M, ϵ) but varying δ. In other words, the function $\nu^\star(\delta) : (0,1) \mapsto \mathbb{R}_+$ is defined for any given (M, ϵ), and we do not associate the change of (M, ϵ) and the change of δ together. To emphasize that we now work with a *fixed* ratio, we shall use δ_0 in place of δ. At this given ratio δ_0, the quantities $\alpha^\star$ and $\nu^\star$ are treated as functions of ϵ.

Proof for part (d) of Theorem 1. When $\epsilon \to 0$, we first make a key observation on the behavior of $\frac{\nu^{\star\prime}(\delta_0)}{M}$, as summarized in the lemma below.

Lemma 3. *In the setting of Theorem 1, given any fixed SNR$= \epsilon M^2$ and $\delta_0 \in (0,1)$, the derivative $\nu^{\star\prime}$ (with respect to δ) obeys*

$$\lim_{\epsilon \to 0} \frac{\nu^{\star\prime}(\delta_0)}{M} > 0. \tag{41}$$

The proof of Lemma 3 contains two main parts, whose details are deferred to Section D.4. First, letting $\alpha_0 \coloneqq -\Phi^{-1}(\delta_0/2)$ for this given δ_0, we establish the following relation

$$\alpha^\star \to \alpha_0; \quad \text{and} \quad \frac{\nu^\star}{M} \to \sqrt{1 - 2\delta_0^{-1}[-\alpha_0\phi(\alpha_0) + (\alpha_0^2+1)\Phi(-\alpha_0)]}, \tag{42}$$

as one takes $\epsilon \to 0$; the details can be found in Section D.4. It is worth noting that both $\alpha^\star$ and $\frac{\nu^\star}{M}$ converge to fixed quantities that are determined only by δ_0 in this limit.

Equipped with these two limiting values, we proceed to consider the numerator and denominator of $\frac{\nu^{\star\prime}}{M}$, with the assistance of the expression (37). In fact, one can pin down the limiting orders of these two parts as follows

$$\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1 \;\to\; 2\delta_0^{-2}\phi(\alpha_0)\left[-2\alpha_0\phi(\alpha_0) + (\alpha_0^2+1)\delta_0\right] - 2\delta_0^{-1}\left[2\phi(\alpha_0) - \alpha_0\delta_0\right], \tag{43a}$$

and

$$M(\nabla_\nu F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\nu F_1) \;\to\; -4\phi(\alpha_0)\sqrt{1-2\delta_0^{-1}[-\alpha_0\phi(\alpha_0)+(\alpha_0^2+1)\Phi(-\alpha_0)]}. \tag{43b}$$

The details can be found in Step 2 in Section D.4.

Putting these together, we can conclude that

$$\lim_{\epsilon\to 0}\frac{\nu^{\star\prime}(\delta_0)}{M} = \frac{2\delta_0^{-2}\left[-2\alpha_0\phi^2(\alpha_0)+(\alpha_0^2+1)\delta_0\phi(\alpha_0)-2\delta_0\phi(\alpha_0)+\alpha_0\delta_0^2\right]}{4\phi(\alpha_0)\sqrt{1-2\delta_0^{-1}[-\alpha_0\phi(\alpha_0)+(\alpha_0^2+1)\Phi(-\alpha_0)]}} < 0,$$

where the last inequality follows due to the fact that $\Phi(-\alpha_0)=\delta_0/2$ and the basic relation

$$\Phi(-\alpha_0) \in \left[\phi(\alpha_0)\left(\frac{1}{\alpha_0}-\frac{1}{\alpha_0^3}\right),\ \phi(\alpha_0)\left(\frac{1}{\alpha_0}-\frac{1}{\alpha_0^3}+\frac{1}{\alpha_0^5}\right)\right].$$

In summary, in view of Lemma 3, we can conclude that there exists $\epsilon^\star>0$, depending only on SNR and δ_0, such that: when $\epsilon<\epsilon^\star$, one has $\nu^{\star\prime}(\delta_0)/M<0$. Translating this back to $\tau^\star = M/\nu^\star$ ensures the existence of an $\epsilon^\star$ such that: when $\epsilon<\epsilon^\star$, one has $\tau^{\star\prime}<0$. We have thus completed the proof of Part (d) of Theorem 1.

Proof of part (c) of Theorem 1. The idea for proving this result is to find $\delta\in(0,1)$ such that the value of $\tau^{\star 2}(\delta)$ is strictly below $\mathsf{Risk}(\mathbf{0})$. Recognizing that $\tau^{\star 2}(\delta)$ decays to $\mathsf{Risk}(\mathbf{0})$ as p/n approaches infinity, there must exist an ascending regime for $\tau^\star$ as a function of p/n.

More concretely, let us view $\nu^\star/M$ as a function of δ within the interval $\delta\in(0,1)$. Rewriting the relation (42) ensures that as $\epsilon\to 0$, one has

$$\frac{1}{\tau^\star(\delta)} = \frac{\nu^\star(\delta)}{M} \to \sqrt{\frac{\alpha\phi(\alpha)-\alpha^2\Phi(-\alpha)}{\Phi(-\alpha)}} =: H(\delta) \qquad \text{for } \alpha := -\Phi^{-1}(\delta/2).$$

It can be easily verified that the function $H(\cdot)$ is a continuous and decreasing function of δ on $(0,1)$. Additionally, direct calculations yield

$$\lim_{\delta\to 1^-} H(\delta) = 0; \qquad \lim_{\delta\to 0^+} H(\delta) = 1. \tag{44}$$

As a result, the continuity of $H(\cdot)$ guarantees that there exists $\delta_{\mathrm{SNR}}>0$ such that

$$H(\delta) := \lim_{\epsilon\to 0}\frac{\nu^\star(\delta)}{M} > \frac{1}{\sqrt{1+\mathrm{SNR}}} = \frac{1}{\sqrt{\mathsf{Risk}(\mathbf{0})}} \in (0,1), \qquad \text{for } \delta<\delta_{\mathrm{SNR}},$$

where we recall $\mathrm{SNR} := \epsilon M^2$. In other words, recognizing the relation $\tau^\star(\delta) := M/\nu^\star(\delta)$, we can show the existence of a regime for $\delta\in(0,1)$ where the ϵ-limit of $\tau^\star(\delta)$ lies below $\sqrt{\mathsf{Risk}(\mathbf{0})}$.

In addition, for any given (ϵ, M), recall that the limiting value (as $\delta\to 0^+$) obeys

$$\lim_{\delta\to 0^+}\frac{\nu^\star(\delta)}{M} = \frac{1}{\sqrt{1+\mathrm{SNR}}}. \tag{45}$$

It further implies that for any fixed $\delta_0<\delta_{\mathrm{SNR}}$, one can find a corresponding ϵ_0 depending on δ_0 such that

$$\tau^\star(\delta_0) < \lim_{\delta\to 0^+}\tau^\star(\delta)$$

holds for every $\epsilon\le\epsilon_0$. Consequently, $\tau^\star(\delta)$ has an ascending phase w.r.t. p/n. Putting the above pieces together establishes the claimed result.

4 Numerical simulations and discussion

This section conducts numerical experiments to confirm the applicability of our results in finite samples and non-Gaussian designs. Along the way, we shall also point out several directions worthy of future investigation.

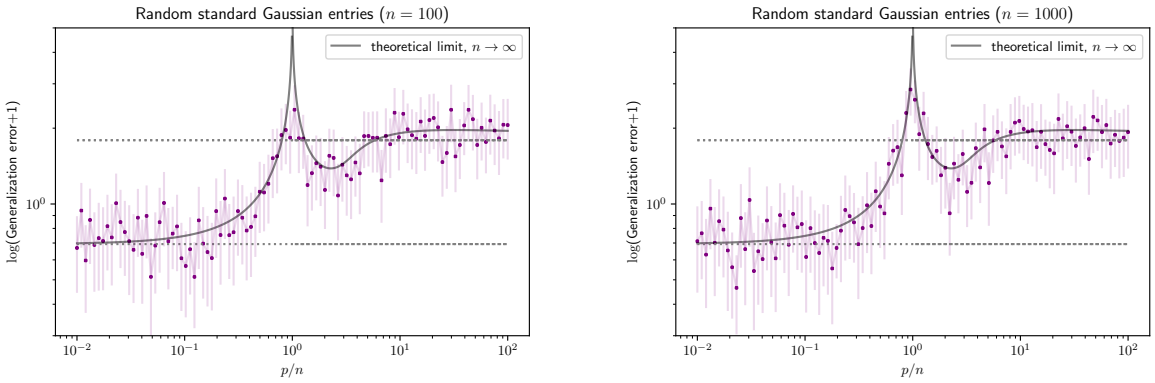

Figure 5. Finite-sample behavior. The data are generated from a linear model (1) under i.i.d. Gaussian design, where SNR= 4 and the sparsity level is $\epsilon = 0.05$. The sample size is set as $n = 100$ in the left figure, and $n = 1000$ in the right figure. The theoretical curve, computed by solving the equations (15), is displayed in solid line, where the limits with $p/n \to 0$ and $p/n \to \infty$ are plotted in dotted lines. Here, both the x-axis and the y-axis are plotted in logarithmic scale. We choose 100 different values of p/n in a way that the $\log(p/n)$'s are uniformly spaced over $[-2, 2]$. For each p/n, we generate a random instance, compute the minimum ℓ_1-norm interpolator and its risk, and repeat this procedure for 30 times. We report the average risk and error bar over 30 independent runs.

Finite-sample behavior. Although the theorems obtained in the paper are asymptotic in nature, our numerical experiments suggest that they are accurate descriptions of the risk behavior even when p and n are on the order of 10s or 100s. As an illustration, we plot in Figure 5 two cases when $n = 100$ and $n = 1000$, respectively, with p/n varying between $[10^{-2},\ 10^2]$. In these plots, the multi-descent phenomenon already manifests itself in the case when $n = 100$.

Beyond Gaussian design. Thus far, our risk characterization focuses on the idealistic case with i.i.d. Gaussian design matrices. There is no shortage of practical scenarios where such distributional assumptions are violated. To see whether our prediction continues to be valid beyond Gaussian design, we carry out several empirical experiments concerning design matrices that are composed of i.i.d. non-Gaussian entries. Figure 6 illustrates two cases where the entries are generated from the Bernoulli distributions and the t-distribution with parameter 3, respectively. Our theoretical risk characterization remains fairly accurate in these numerical experiments. This is perhaps not unexpected, due to a *universality* phemoneon that has been justified in multiple other problems with i.i.d. random design (see, e.g., Bayati et al. (2015); Oymak and Tropp (2018); Montanari and Nguyen (2017); Chen and Lam (2021)). These predictions might, however, be completely off when the covariates are correlated, meaning that the covariance structure of the design matrix plays a pivotal role in determining the shape of the risk curves. Leveraging the current effort towards understanding Lasso under correlated designs (Celentano et al., 2020), we conjecture that the risk of the interpolator is dictated by a more complicated nonlinear system of equations that reflects the covariance structure. Given that the main message of this paper is to verify the existence of a multiple-descent phenomenon, we leave these more general cases to future investigation.

Distributional characterization. We perform another series of numerical experiments about the minimum ℓ_1-norm interpolators $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ under i.i.d. Gaussian design, and report in Figure 7 *(i)* the empirical distribution of its p coordinates over 30 independent runs, and *(ii)* the empirical distribution of the corresponding $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ coordinates when the underlying $\theta_i^\star$ is zero (resp. non-zero). As can be seen from the plots, the estimates are close to being unbiased, with the estimates for non-zero entries exhibiting a higher level of uncertainty than the zero entries. However, how to develop a distributional theory remains unclear. A recent line of works (Bellec and Zhang, 2019; Miolane and Montanari, 2018; Celentano et al., 2020) established distributional guarantees for a debiased Lasso estimator with positive regularization (so that the estimates after de-biasing exhibit Gaussian distributions). We conjecture that the analysis framework (via the convex Gaussian min-max theorem) developed in Miolane and Montanari (2018); Celentano et al. (2020) might be useful in establishing

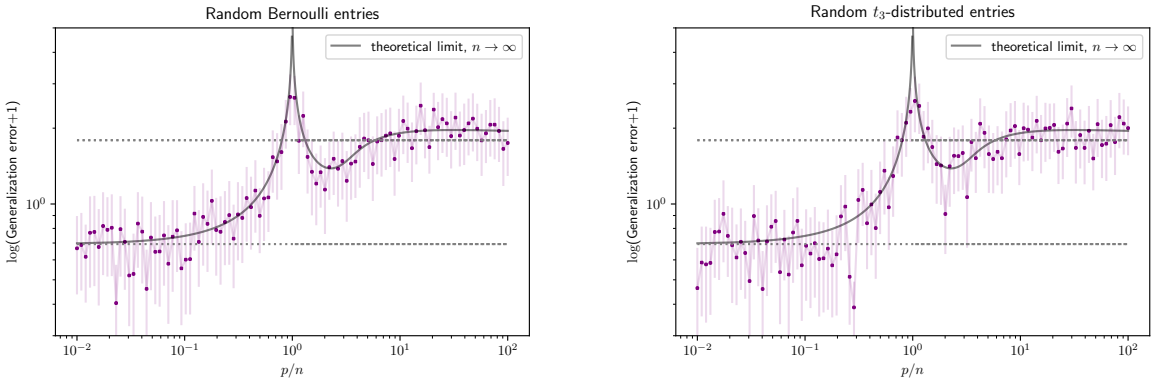

Figure 6. Experiments for non-Gaussian designs. In these plots, the sample size is fixed as $n = 100$, and the data is drawn from a linear model (1) with SNR$= 4$ and sparsity level $\epsilon = 0.05$. The entries of the design matrix $\sqrt{n}\boldsymbol{X}$ are i.i.d. sampled from the Bernoulli(0.5) distribution for the left plot, and from $t(3)/\sqrt{3}$ distribution for the right plot (where the $1/\sqrt{3}$ is introduced to make the variance equals to 1). The other experiment settings are the same with Figure 5.

a fine-grained finite-sample distributional characterization for the interpolators of interest.

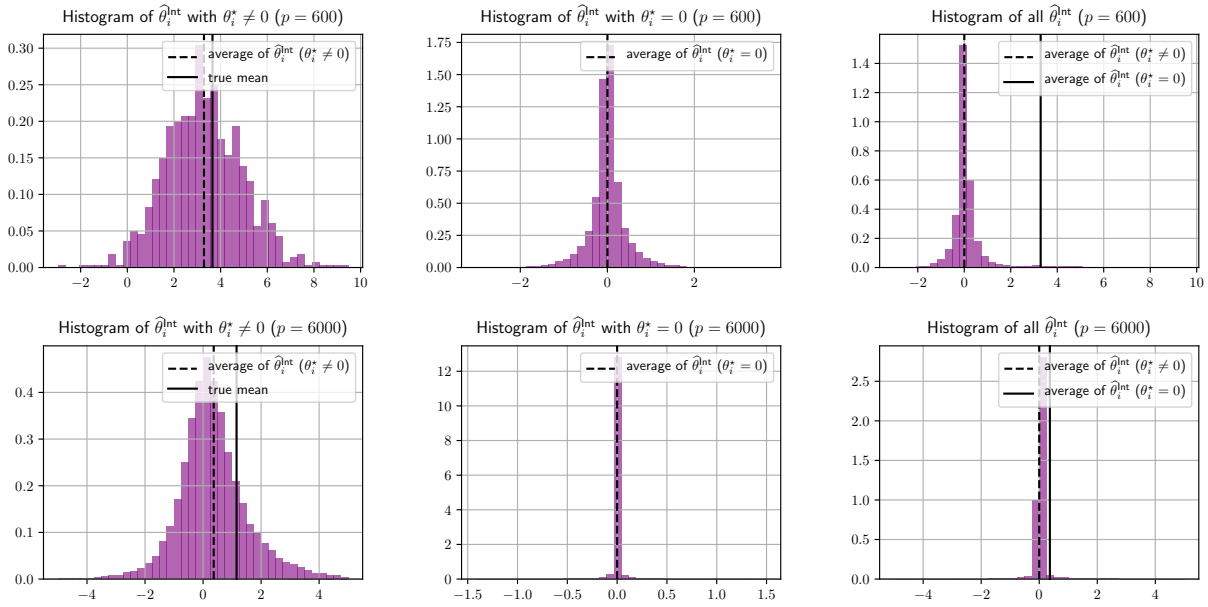

Figure 7. Empirical distribution for coordinates of $\widehat{\theta}_i^{\mathsf{Int}}$. Here, we fix the sample size $n = 100$, and generate data from the linear model (1) with i.i.d. Gaussian design, where SNR$= 4$ and sparsity level $\epsilon = 0.05$. The other experiment settings are the same with Figure 5. We collect the empirical distribution of $\widehat{\theta}_i^{\mathsf{Int}}$'s coordinates (corresponding to those i such that $\theta_i^\star \neq 0$ / $\theta_i^\star = 0$ / for every i, respectively) across all repeats, and generate their histograms. In the top row of the plots, set $p = 600$, and in the bottom row, set $p = 6000$. All the results reported are based on 30 random trials. The empirical averages and the ground truth $\theta_i^\star$ values are marked in the dotted vertical line and the solid vertical line respectively.

Acknowledgment

The authors would like to thank Linjun Zhang for discussing this open problem with Y. Wei when she was visiting the statistics department at Rutgers University in 2020. This work was partially supported by the NSF grants DMS 2147546/2015447 and CCF 2106778. Part of this work was done while Y. Li and Y. Wei

were visiting the Simons Institute for the Theory of Computing.

References

Adlam, B. and Pennington, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bayati, M., Lelarge, M., and Montanari, A. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822.

Bayati, M. and Montanari, A. (2011a). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

Bayati, M. and Montanari, A. (2011b). The LASSO risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.

Belkin, M., Ma, S., and Mandal, S. (2018). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR.

Bellec, P. C. and Zhang, C.-H. (2019). Second order Poincaré inequalities and de-biasing arbitrary convex regularizers when $p/n \to \gamma$. *arXiv preprint arXiv:1912.11943*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.

Bu, Z., Klusowski, J. M., Rush, C., and Su, W. J. (2020). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on Information Theory*, 52(12):5406–5425.

Celentano, M., Fan, Z., and Mei, S. (2021). Local convexity of the TAP free energy and AMP convergence for Z2-synchronization. *arXiv preprint arXiv:2106.11428*.

Celentano, M., Montanari, A., and Wei, Y. (2020). The Lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*.

Chen, L., Min, Y., Belkin, M., and Karbasi, A. (2020). Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*.

Chen, S. and Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159.

Chen, W.-K. and Lam, W.-K. (2021). Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44.

Chen, Z. and Dongarra, J. J. (2005). Condition numbers of gaussian random matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(3):603–620.

Chinot, G., Löffler, M., and van de Geer, S. (2020). On the robustness of minimum-norm interpolators. *arXiv preprint arXiv:2012.00807*.

d'Ascoli, S., Sagun, L., and Biroli, G. (2020). Triple descent and the two kinds of overfitting: Where & why do they appear? *arXiv preprint arXiv:2006.03509*.

Deshpande, Y., Abbe, E., and Montanari, A. (2015). Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

Donoho, D. and Tanner, J. (2009). Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.

Donoho, D. L., Elad, M., and Temlyakov, V. N. (2005). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18.

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

Donoho, D. L., Maleki, A., and Montanari, A. (2010). Message passing algorithms for compressed sensing: II. Analysis and validation. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5. IEEE.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

Fan, Z. (2020). Approximate message passing algorithms for rotationally invariant matrices. *accepted to the Annals of Statistics*.

Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. (2021). A unifying tutorial on approximate message passing. *arXiv preprint arXiv:2105.02180*.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2018). Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Javanmard, A. and Montanari, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144.

Ju, P., Lin, X., and Liu, J. (2020). Overfitting can be harmless for basis pursuit, but only to a degree. *arXiv preprint arXiv:2002.00492*.

Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347.

Liang, T., Rakhlin, A., and Zhai, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR.

Liang, T. and Sur, P. (2020). A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*.

Ma, J., Xu, J., and Maleki, A. (2018). Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization. *arXiv preprint arXiv:1801.01170*.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*.

Miolane, L. and Montanari, A. (2018). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*.

Mitra, P. P. (2019). Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*.

Montanari, A. and Nguyen, P.-M. (2017). Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE.

Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*.

Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. (2020). Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*.

Neyshabur, B., Tomioka, R., and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Oymak, S., Thrampoulidis, C., and Hassibi, B. (2013). The squared-error of generalized Lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE.

Oymak, S. and Tropp, J. A. (2018). Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR.

Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973.

Rush, C. and Venkataramanan, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286.

Stojnic, M. (2013). A framework to characterize performance of LASSO algorithms. *arXiv preprint arXiv:1303.7291*.

Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of statistics*, pages 2133–2150.

Su, W. and Candes, E. (2016). Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Viering, T., Mey, A., and Loog, M. (2019). Open problem: Monotonicity of learning. In *Conference on Learning Theory*, pages 3198–3201. PMLR.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wei, Y., Yang, F., and Wainwright, M. J. (2019). Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 65(10):6685–6703.

Wojtaszczyk, P. (2010). Stability and instance optimality for gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10(1):1–13.

Wyner, A. J., Olson, M., Bleich, J., and Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zhong, X., Wang, T., and Fan, Z. (2021). Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *arXiv preprint arXiv:2110.02318*.

APPENDIX

A Proof of Proposition 1

In what follows, we intend to establish the two claims separately.

Case I: $n/p > 1$. In this part, we aim to prove that, for any $\delta > 1$, the Lasso risk converges to the risk of the least-square estimator — denoted by $\widehat{\boldsymbol{\theta}}^{\mathsf{LS}}$ — as $\lambda \to 0$. To begin with, the risk of $\widehat{\boldsymbol{\theta}}^{\mathsf{LS}}$ can be characterized using standard random matrix theory results; see, for example, Hastie et al. (2019, Theorem 1). As $p \to \infty$, one has

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{LS}}) = \sigma^2 + \frac{1}{n}\mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^{\mathsf{LS}} - \boldsymbol{\theta}^\star\|_2^2\big] \to \sigma^2 \frac{\delta}{\delta - 1}. \tag{46}$$

In view of the KKT condition for the corresponding loss functions, we can see that the Lasso and the least-square estimator obey

$$\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^\star = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{X}^\top \boldsymbol{z} - \lambda \boldsymbol{s}) \qquad \text{and} \qquad \widehat{\boldsymbol{\theta}}^{\mathsf{LS}} - \boldsymbol{\theta}^\star = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{z}.$$

Here, $\boldsymbol{s} = [s_j]_{1\le j\le p}$ denotes the sub-gradient of the ℓ_1 norm at point $\widehat{\boldsymbol{\theta}}_\lambda$, which obeys $s_j \in [-1, 1]$. Thus, the risk of the Lasso satisfies

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^\star\|_2^2\big] = \mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^{\mathsf{LS}} - \boldsymbol{\theta}^\star\|_2^2\big] - 2\lambda\mathbb{E}\big[\langle (\boldsymbol{X}^\top \boldsymbol{X})^{-2}\boldsymbol{X}^\top \boldsymbol{z},\, \boldsymbol{s}\rangle\big] + \lambda^2 \mathbb{E}\big[\|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{s}\|_2^2\big],$$

which combined with the Cauchy-Schwarz inequality further leads to

$$\begin{aligned}
&\left|\frac{1}{n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^\star\|_2^2] - \frac{1}{n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\mathsf{LS}} - \boldsymbol{\theta}^\star\|_2^2]\right| \\
&\le 2\lambda\sqrt{\frac{1}{n}\mathbb{E}\left[\|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{z}\|_2^2\right]}\sqrt{\frac{1}{n}\mathbb{E}\big[\|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\|_2^2\|\boldsymbol{s}\|_2^2\big]} + \lambda^2 \frac{1}{n}\mathbb{E}\left[\|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\|_2^2\|\boldsymbol{s}\|_2^2\right] \\
&\le 2\lambda\sqrt{\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{LS}}) - \sigma^2}\sqrt{\mathbb{E}\left[\frac{1}{\delta}\sigma_{\min}^{-4}(\boldsymbol{X})\right]} + \frac{\lambda^2}{\delta}\mathbb{E}\left[\sigma_{\min}^{-4}(\boldsymbol{X})\right].
\end{aligned} \tag{47}$$

Now it is sufficient to control the two terms on the right-hand side above, and show that both terms converge to 0 when $p \to \infty$. First, it has been shown in the proof of Chen and Dongarra (2005, Lemma 4.1) that

$$\mathbb{P}\left(\sigma_{\min}(\boldsymbol{X}) \le \frac{\sqrt{n}}{x^2}\right) < \frac{n^{n-p+1}}{(n-p+1)!}\frac{1}{x^{n-p+1}} \le \left(\frac{e}{x}\right)^{n-p+1}$$

for any $x > 0$, where the last inequality comes from the well-known Stirling inequality $\sqrt{2\pi}m^{m+\frac{1}{2}}e^{-m} \le m!$ (Robbins, 1955). Consequently,

$$\begin{aligned}
\mathbb{E}\left[\sigma_{\min}^{-4}(\boldsymbol{X})\right] &\le \left(\frac{2}{1-1/\sqrt{\delta}}\right)^4 \mathbb{P}\left\{\sigma_{\min}^{-1}(\boldsymbol{X}) \le \frac{2}{1-1/\sqrt{\delta}}\right\} + \int_{\left(\frac{2}{1-1/\sqrt{\delta}}\right)^4}^{\infty} \mathbb{P}\left\{\sigma_{\min}^{-4}(\boldsymbol{X}) > z\right\} \mathrm{d}z \\
&\le 2\left(\frac{2}{1-1/\sqrt{\delta}}\right)^4 + \int_{\left(\frac{2}{1-1/\sqrt{\delta}}\right)^4}^{\infty} \mathbb{P}\left\{\sigma_{\min}(\boldsymbol{X}) < \frac{1}{z^{1/4}}\right\} \mathrm{d}z \\
&\le 2\left(\frac{2}{1-1/\sqrt{\delta}}\right)^4 + \int_{\left(\frac{2}{1-1/\sqrt{\delta}}\right)^4}^{\infty} \left(\frac{e}{z^{1/8}n^{1/4}}\right)^{n-p+1} \mathrm{d}z \\
&\le 4\left(\frac{2}{1-1/\sqrt{\delta}}\right)^4
\end{aligned}$$

for sufficiently large n. In addition, by virtue of (46), it is guaranteed that

$$\sqrt{\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{LS}}) - \sigma^2} \to \sqrt{\frac{\sigma^2}{\delta - 1}}.$$

Substitution into (47) yields

$$\lim_{\lambda \to 0} \lim_{n \to \infty} \frac{1}{n} \mathbb{E}[\|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^\star\|_2^2] = \lim_{n \to \infty} \frac{1}{n} \mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^{\mathsf{LS}} - \boldsymbol{\theta}^\star\|_2^2\big]$$

for any $0 < \delta < 1$ that is strictly bounded away from 1.

Case II: $n/p < 1$**.** When $\delta < 1$, our goal is to demonstrate that

$$\lim_{\lambda \to 0} \lim_{p \to \infty} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda) = \lim_{p \to \infty} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}).$$

To this end, let us first state one known result about $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda)$. Specifically, the lemma below associates the Lasso risk with the solution to the system of equations (19a) and (19b).

Lemma A.1 (Corollary 1.6 in Bayati and Montanari (2011b))**.** *The system of equations* (19) *admits one unique solution pair* $(\tau^\star(\lambda), \alpha^\star(\lambda))$*. With the Lasso problem formulated in* (18)*, it holds that*

$$\lim_{p \to \infty} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}_\lambda) = \big(\tau^\star(\lambda)\big)^2,$$

as long as $\mathbb{P}(\boldsymbol{\theta}^\star \neq \mathbf{0}) > 0$*.*

In view of Theorem 2, it is guaranteed that

$$\lim_{p \to \infty} \mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}) = \tau^{\star 2}.$$

Therefore, to obtain the desired conclusion, it suffices to show that $\lim_{\lambda \to 0} \tau^\star(\lambda) = \tau^\star$, where $\tau^\star(\lambda)$, and $\tau^\star$ correspond to the solution to a different set of equations respectively. Equivalently, for any converging sequence $\{\lambda_t\}_{t=1}^{+\infty}$ with $\lambda_t > 0$ and $\lambda_t \to 0$, denote the corresponding $(\tau^\star(\lambda_t), \alpha^\star(\lambda_t))$ sequence as $\{(\tau_t^\star, \alpha_t^\star)\}$. We now aim to show that the $\lim_{t \to \infty} \tau_t^\star = \tau^\star$.

In order to achieve this goal, we make two useful observations. First, as will be demonstrated in Lemma B.1(1), we know that as $\lambda_t \to 0$, $\{\alpha_t^\star\}$ is a non-increasing sequence and is lower bounded by $\alpha_{\min}(\delta)$. Therefore, $\{\alpha_t^\star\}$ has a finite and positive limit; we shall denote this limit by $\alpha_\infty^\star$. In addition, applying Lemma B.1(2) ensures that $\{\tau_t^\star = \tau_*(\alpha_t^\star)\}$ converges; we shall denote the limiting value by $\tau_\infty^\star$. Consequently, taking $t \to \infty$ on both sides of

$$\lambda_t = \alpha_t^\star \tau_t^\star \left(1 - \frac{1}{\delta} \mathbb{E}\left[\eta'(\Theta + \tau_t^\star Z; \alpha_t^\star \tau_t^\star)\right]\right),$$

(note that the right-hand side is continuously differentiable with respect to both parameters) leads to the observation that $(\alpha_\infty^\star, \tau_\infty^\star)$ yields $\delta = \mathbb{P}\left(|\Theta + \tau_\infty^\star Z| > \alpha_\infty^\star \tau_\infty^\star\right)$, thus solving the equation (15b). Similarly, one can show that $(\alpha_\infty^\star, \tau_\infty^\star)$ solves the equation (15a).

Putting these pieces together and using the uniqueness of the solution to (15a) and (15b), we arrive that $\lim_{t \to \infty} \tau_t^\star = \tau_\infty^\star = \tau^\star$. We have thus established the advertised property.

B Properties of the state evolution parameters

In this section, we collect some results about the state evolution parameters $\{\alpha_t^\star, \tau_t^\star, \tau_t\}_{t=1}^\infty$. We remind the readers that τ_t is the state evolution in the t-th iteration, while $(\alpha_t^\star, \tau_t^\star) = (\alpha^\star(\lambda_t), \tau^\star(\lambda_t))$ represents the fixed point of the state evolution recursion with λ_t.

B.1 Main results

We first make note of several useful results about the solutions to the equations (19a) and (19b), which have been proved in Bayati and Montanari (2011b). Before proceeding, first recall the following mapping $\mathsf{F}(\tau^2, \zeta)$ previously introduced in (21b):

$$\mathsf{F}(\tau^2, \zeta) = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{\left[\eta(\Theta + \tau Z; \zeta) - \Theta\right]^2\right\},$$

where $Z \sim \mathcal{N}(0,1)$ is independent of Θ. We also recall that $\alpha_{\min} = \alpha_{\min}(\delta)$ corresponds to the non-negative solution of the equation

$$(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) = \frac{\delta}{2}.$$

We now record the following properties.

Lemma B.1 (Proposition 1.3, Proposition 1.4, Corollary 1.7 in Bayati and Montanari (2011b))**.** *The solution to the equations (19a) and (19b) obeys*

1. *For any $\alpha > \alpha_{\min}(\delta)$, the first equation $\tau^2 = \mathsf{F}(\tau^2, \alpha\tau)$ admits a unique solution; denote this solution as $\tau^\star = \tau^\star(\alpha)$. Additionally we know that $\alpha \mapsto \tau^\star(\alpha)$ is continuously differentiable on $(\alpha_{\min}(\delta), +\infty)$.*
2. *For any $\lambda > 0$, there exists one unique $\alpha > 0$ satisfying (19a) and (19b) with $\alpha > \alpha_{\min}(\delta)$, and the mapping from $\lambda > 0$ to α is continuous, differentiable and non-decreasing. Its inverse mapping*

$$\lambda(\alpha) \coloneqq \alpha\tau^\star(\alpha)\left[1 - \frac{1}{\delta}\mathbb{E}\left\{\eta'(\Theta + \tau^\star(\alpha)Z; \alpha\tau^\star(\alpha))\right\}\right] \tag{48}$$

   is continuous and differentiable on $(\alpha_{\min}(\delta), +\infty)$, with $\lambda(\alpha_{\min}(\delta)+) = -\infty$ and $\lim_{\alpha\to\infty}\lambda(\alpha) = +\infty$.
3. *Combining 1 and 2, we can see that for any $\lambda > 0$, there is a unique pair $(\alpha^\star(\lambda), \tau^\star(\lambda))$ that solves (19a) and (19b).*

Through careful investigations about the constructed mappings $\tau^\star(\alpha)$ and $\alpha^\star(\lambda)$ for every $\lambda > 0$, we establish the existence and uniqueness of the solution to the equations (15a) and (15b). The proof of this result is provided in Section B.2.

Proposition B.1 (Uniqueness of the solution)**.** *When $\mathbb{P}(\Theta \neq 0) > 0$, there exists one unique pair $(\alpha^\star, \tau^\star)$ with $\alpha^\star > \alpha_{\min}(\delta)$ that satisfies (15a) and (15b).*

It turns out that both $\tau^\star(\alpha)$ and $\alpha^\star(\lambda)$ are Lipschitz functions of λ, which can be rigorized in the following proposition. The proof of this result can be found in Section B.2.

Proposition B.2 (Convergence of $\alpha_t^\star$ and $\tau_t^\star$)**.** *The following properties hold for the solutions to the equation set (26) as $t \to \infty$.*

1. *$\alpha_t^\star \to \alpha^\star$, $\tau_t^\star \to \tau^\star$; the sequence $\{\alpha_t^\star\}$ increases monotonically with t when t is large enough.*
2. *The function $\alpha^\star(\lambda)$ is continuously differentiable w.r.t. λ, and there exist some constants c and C determined by Θ, σ and δ, such that $c \le \alpha^{\star\prime}(0) \le C$; as such, we can find some L_α determined by Θ, σ and δ, such that*

$$|\alpha_t^\star - \alpha_{t+1}^\star| \le L_\alpha|\lambda_t - \lambda_{t+1}|.$$

3. *The function $\tau^\star(\lambda)$ is Lipschitz w.r.t. λ for some L_τ determined by Θ, σ and δ; in particular,*

$$|\tau_t^\star - \tau_{t+1}^\star| \le L_\tau|\lambda_t - \lambda_{t+1}|.$$

As a direct consequence of the third claim above, for every $t \ge 1$, one has

$$\tau_t^\star \le \tau^\star + L_\tau \max_i \lambda_i =: \tau_{\max}^\star; \quad \alpha_t^\star \le \alpha^\star + L_\alpha \max_i \lambda_i =: \alpha_{\max}^\star.$$

Finally, we can demonstrate that: the state evolution sequence τ_t (defined in (24)) approaches the solution of the equations (15a) and (15b) as $t \to \infty$. The proof is deferred to Section B.3.

Proposition B.3 (Convergence of the state evolution)**.** *The state evolution sequence obeys $\tau_t \to \tau^\star$ as $t \to \infty$.*

B.2 Proof of Proposition B.1 and Proposition B.2

To begin with, we introduce the following result on the derivatives of $\lambda'(\alpha)$ and $\tau^{\star\prime}(\alpha)$, whose proof is provided in Section B.2.1.

Lemma B.2. *For any $\alpha_0 > \alpha_{\min}(\delta)$ with $\lambda(\alpha_0) = 0$ (cf. (48)), we have*

$$0 < C_1 < \lambda'(\alpha_0) < C_2; \tag{49}$$

$$\tau^{\star\prime}(\alpha_0) \le C_3. \tag{50}$$

Here, C_1, C_2, C_3 are constants that depend only on α_0, δ and Θ.

We are ready to prove Proposition B.1 and Proposition B.2 with the assistance of Lemma B.2.

Proof of Proposition B.1. Consider the function $\lambda(\alpha)$ defined in (48). Suppose there exist two different $\alpha_0 \neq \alpha_0'$ where $\lambda(\alpha_0) = \lambda(\alpha_0') = 0$; by Lemma B.2, we know $\lambda'(\alpha_0)$ and $\lambda'(\alpha_0')$ are both positive and bounded away from 0. Thus, in view of the continuity of $\lambda(\cdot)$, there must exist some $\lambda_0 > 0$ such that $\lambda(\alpha) = \lambda_0$ has at least two solutions. This, however, contradicts Lemma B.1(2).

Proof of Proposition B.2. We consider each claim separately. For the first claim, by virtue of the second statement in Lemma B.1, the mapping $\alpha \mapsto \lambda(\alpha)$ is continuous and differentiable on $(\alpha_{\min}, +\infty)$, and for any $\lambda \ge 0$, the solution of $\lambda(\alpha) = \lambda$ exists and is unique. As such, the inverse mapping $\lambda \mapsto \alpha^\star(\lambda)$ is well-defined on $\lambda \ge 0$ and is continuous. Then we can see that $\alpha_t^\star \to \alpha^\star$. Moreover, recognizing that $\tau_t^\star = \tau^\star(\alpha_t^\star)$ and $\tau^\star = \tau^\star(\alpha^\star)$ and using the continuously differentiable mapping $\alpha \mapsto \tau^\star(\alpha)$ defined in Lemma B.1(1), we conclude that $\tau_t^\star \to \tau^\star$. Lastly, in light of inequality (49), when $\lambda_t \neq \lambda_{t+1}$ we have

$$\frac{\alpha_t^\star - \alpha_{t+1}^\star}{\lambda_t - \lambda_{t+1}} \to \alpha^{\star\prime}(0) \ge \frac{1}{C_1}.$$

We can thus conclude that $\alpha_t^\star$ is monotonously increasing when t is large enough.

We now turn to the second claim. As ensured by Proposition B.1, there exists one unique solution to $\lambda(\alpha_0)$; therefore, the expression (49) translates to

$$0 < C_1 < \lambda'(\alpha^\star) < C_2,$$

or equivalently, $C_2^{-1} < \alpha^{\star\prime}(0) < C_1^{-1}$. We have thus finished the proof of the second claim.

The third claim follows directly from the second claim and the bound (50), and the proof of Proposition B.1 is completed. Finally, we make the remark that L_τ and L_α depend on C_1, C_2 and C_3 from Lemma B.2; from Proposition B.1, we know that $\alpha_0 = \alpha^\star$ is the unique value of α_0 satisfying Proposition B.1. Therefore, the expressions (55) (56)(58) and (57) with $\alpha_0 = \alpha^\star$ and $\tau^\star(\alpha_0) = \tau^\star$ give us the explicit from of L_τ and L_α in terms of $\alpha^\star$ and $\tau^\star$.

B.2.1 Proof of Lemma B.2

We first make note of an inequality proved in (Miolane and Montanari, 2018, Lemma A.5) as follows:

$$\tau^{\star\prime}(\alpha_0) \le (\alpha_0 + 1)\frac{\tau^{\star 3}(\alpha_0)}{\delta\sigma^2} =: C_3. \tag{51}$$

which validates the inequality (50).

It then suffices to establish the first inequality. Towards this, let us first derive the explicit expressions for the derivative, and prove the upper and lower bounds. For any $\alpha > \alpha_{\min}(\delta)$, define $\Xi := \Theta/\tau^\star(\alpha)$, and the following quantities:

$$E_1 := \mathbb{E}\left[\Phi(-\Xi - \alpha) + \Phi(\Xi - \alpha)\right], \quad E_2 := \mathbb{E}\left[\Xi\phi(-\Xi - \alpha) - \Xi\phi(\Xi - \alpha)\right],$$
$$E_3 := \mathbb{E}\left[\phi(-\Xi - \alpha) + \phi(\Xi - \alpha)\right], \quad E_4 := \mathbb{E}\left[\Xi^2\left[\Phi(\alpha - \Xi) - \Phi(-\alpha - \Xi)\right]\right].$$

It is easily seen that E_1, E_2, E_3 and E_4 are continuous and differentiable with respect to α.

Explicit expression for the derivatives at $\alpha = \alpha_0$. Let us first derive $\tau^{\star\prime}(\alpha)$ with $\tau^\star(\alpha)$ defined in Lemma B.1(1). Recall the definition $\mathsf{F}(\tau^2, \zeta) = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta(\Theta + \tau Z; \zeta) - \Theta]^2\right\}$; direct calculations yield

$$\begin{aligned}
\frac{\partial \mathsf{F}}{\partial \alpha}(\tau^2, \tau\alpha) =& \frac{2\tau^\star(\alpha)^2}{\delta}\left\{\alpha\mathbb{E}\left[\Phi(\Xi - \alpha) + \Phi(-\Xi - \alpha)\right] - \mathbb{E}\left[\phi(\Xi - \alpha) + \phi(-\Xi - \alpha)\right]\right\} = \frac{2\tau^\star(\alpha)^2}{\delta}\left[\alpha E_1 - E_3\right];\\
\frac{\partial \mathsf{F}}{\partial \tau}(\tau^2, \tau\alpha) =& \frac{2\tau^\star(\alpha)}{\delta}\left\{(1+\alpha^2)\mathbb{E}\left[\Phi(\Xi - \alpha) + \Phi(-\Xi - \alpha)\right] - \mathbb{E}\left[(\Xi + \alpha)\phi(\Xi - \alpha) - (\Xi - \alpha)\phi(-\Xi - \alpha)\right]\right\}\\
=& \frac{2\tau^\star(\alpha)}{\delta}\left[(1+\alpha^2)E_1 + E_2 - \alpha E_3\right].
\end{aligned}$$

Invoking the implicit function theorem, we can guarantee that

$$\tau^{\star\prime}(\alpha) = \frac{\partial_\alpha \mathsf{F}(\tau^2, \alpha\tau)}{2\tau - \partial_\tau \mathsf{F}(\tau^2, \alpha\tau)} = -\frac{\tau^\star(\alpha)\left[\alpha E_1 - E_3\right]}{(1+\alpha^2)E_1 + E_2 - \alpha E_3 - \delta}.$$

For α_0 with $\lambda(\alpha_0) = 0$, we know $E_1 = \delta$, and then

$$\tau^{\star\prime}(\alpha_0) = -\frac{\tau^\star(\alpha_0)\left[\alpha_0\delta - E_3\right]}{\alpha_0^2\delta + E_2 - \alpha_0 E_3}.$$

In terms of $\lambda'(\alpha)$, one has

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{E}\left[\eta'(\Theta + \tau_\star(\alpha)Z; \tau^\star(\alpha)\alpha)\right] &= \mathbb{E}\left[(-1 + \Xi\frac{\tau^{\star\prime}(\alpha)}{\tau_\star(\alpha)})\phi(-\Xi - \alpha) + (-1 - \Xi\frac{\tau^{\star\prime}(\alpha)}{\tau_\star(\alpha)})\phi(\Xi - \alpha)\right]\\
&= -\frac{\alpha\left[\alpha E_1 E_3 - E_3^2 + E_1 E_2\right] + E_3(E_1 - \delta)}{(1+\alpha^2)E_1 + E_2 - \alpha E_3 - \delta}.
\end{aligned}$$

Finally, we are ready to calculate the derivative $\lambda'(\alpha)$ at $\alpha = \alpha_0$. By expression (19), and noticing that $\lambda(\alpha_0) = 0$ and $E_1 = \delta$ in this case, we calculate

$$\begin{aligned}
\lambda'(\alpha_0) &= \frac{\alpha_0\tau^{\star\prime}(\alpha_0) + \tau^\star(\alpha_0)}{\alpha_0\tau^\star(\alpha_0)}\lambda(\alpha_0) - \frac{\alpha_0\tau^\star(\alpha_0)}{\delta}\frac{d}{d\alpha}\mathbb{E}\left[\eta'(\Theta + \tau^\star(\alpha)Z; \tau^\star(\alpha)\alpha)\right]\Big|_{\alpha=\alpha_0}\\
&= \frac{\alpha_0^2\tau^\star(\alpha_0)}{\delta}\cdot\frac{\delta\alpha_0 E_3 - E_3^2 + \delta E_2}{\delta\alpha_0^2 + E_2 - \alpha_0 E_3}.
\end{aligned} \tag{52}$$

Bounding the derivatives. To establish the inequality (49), it is sufficient for us to control the following quantities

$$\delta\alpha_0 E_3 - E_3^2 + \delta E_2 \quad \text{and} \quad \delta\alpha_0^2 + E_2 - \alpha_0 E_3,$$

respectively. Let us first express the function $\mathsf{F}(\tau^2, \tau\alpha)$ in an explicit fashion. For every fixed ξ, we have

$$\begin{aligned}
\mathbb{E}\left[\eta(\xi + Z; \alpha) - \xi\right]^2 =& (-\alpha - \xi)\phi(\alpha - \xi) + (\alpha^2 + 1)\Phi(-\alpha + \xi) + (-\alpha + \xi)\phi(\alpha + \xi) + (\alpha^2 + 1)\Phi(-\alpha - \xi)\\
&+ \xi^2\left[\Phi(\alpha - \xi) - \Phi(-\alpha - \xi)\right].
\end{aligned}$$

Taking expectation with respect to $\xi = \Theta/\tau^\star(\alpha)$ on both sides, we arrive at

$$\tau^\star(\alpha)^2 = \mathsf{F}(\tau^\star(\alpha)^2, \tau^\star(\alpha)\alpha) = \sigma^2 + \frac{\tau^\star(\alpha)^2}{\delta}\left[(\alpha^2 + 1)E_1 + E_2 - \alpha E_3 + E_4\right]. \tag{53}$$

For $\alpha_0 > \alpha_{\min}(\delta) > 0$ with $\lambda(\alpha_0) = 0$, one has $E_1 = \delta$ and $E_4 \geq 0$. As a result, the equation (53) leads to

$$\alpha_0^2\delta + E_2 - \alpha E_3 \leq -\sigma^2\delta\tau^{\star -2}(\alpha_0) < 0. \tag{54}$$

Next we control the quantity $\delta\alpha_0 E_3 - E_3^2 + \delta E_2$ by looking at two cases separately. Observing that $\alpha_0^2\delta + E_2 - \alpha E_3 \leq -\sigma^2\delta\tau^{\star -2}(\alpha_0)$, $E_3 \geq 0$ and $E_2 \leq 0$, we know

$$\delta\alpha_0 E_3 - E_3^2 + \delta E_2 < E_3\left[-\frac{\sigma^2\delta}{\tau^{\star 2}(\alpha_0)\alpha_0} - \frac{E_2}{\alpha_0}\right] + \delta E_2 = -\frac{\sigma^2\delta}{\tau^{\star 2}(\alpha_0)\alpha_0}E_3 + \frac{\alpha_0\delta - E_3}{\alpha_0}E_2.$$

Then we have

$$\delta\alpha_0 E_3 - E_3^2 + \delta E_2 \leq \begin{cases} -\dfrac{\sigma^2\delta}{\tau^{\star 2}(\alpha_0)\alpha_0} E_3, & \alpha_0\delta - E_3 > 0; \\ \delta E_2, & \alpha_0\delta - E_3 \leq 0. \end{cases}$$

Taking the above properties collectively with (52), we now move on to prove the conclusion in (49) for two cases respectively, namely,

$$\lambda'(\alpha_0) \begin{cases} \geq \dfrac{\sigma^2\alpha_0}{\tau^\star(\alpha_0)} \dfrac{E_3}{|E_2| - \alpha_0\left[\delta\alpha_0 - E_3\right]} \geq \dfrac{\sigma^2\alpha_0}{\tau^\star(\alpha_0)} \dfrac{E_3}{|E_2|}, & \alpha_0\delta - E_3 > 0; \\ = \dfrac{\alpha_0^2\tau^\star(\alpha_0)}{\delta} \dfrac{E_3\left[\alpha_0\delta - E_3\right] + \delta E_2}{\alpha_0\left[\alpha_0\delta - E_3\right] + E_2} \geq \dfrac{\alpha_0^2\tau^\star(\alpha_0)}{\delta} \min\left\{\delta, \dfrac{E_3}{\alpha_0}\right\} = \min\left\{\alpha_0^2\tau^\star(\alpha_0), \dfrac{\alpha_0\tau^\star(\alpha_0)E_3}{\delta}\right\}, & \alpha_0\delta - E_3 \leq 0. \end{cases} \tag{55}$$

For the upper bound of $\lambda'(\alpha_0)$, we can directly verify that

$$\delta\alpha_0 E_3 - E_3^2 + \delta E_2 \geq \delta E_2 - 4,$$

which combined with the expression (54) leads to

$$\lambda'(\alpha_0) \leq \frac{\alpha_0^2\tau^\star(\alpha_0)}{\delta} \frac{\delta|E_2| + 4}{\sigma^2\delta\tau^{\star -2}(\alpha_0)}. \tag{56}$$

From our assumption $\mathbb{E}[\Theta^2] < \infty$, we can find M, such that

$$\mathbb{E}[\Theta^2 \mathbf{1}\left\{|\Theta| \leq M\right\}] \geq \mathbb{E}[\Theta^2]/2.$$

Then we know that $\mathbb{P}(|\Theta| \leq M) \geq \mathbb{E}[\Theta^2/M^2 \mathbf{1}\left\{|\Theta| \leq M\right\}] \geq \mathbb{E}[\Theta^2]/(2M^2)$. Combining with the definition of E_3 yields

$$E_3 \geq \mathbb{E}\left[\phi(|\Xi| - \alpha_0)\right] \geq \mathbb{E}\left[\phi\left(\frac{|\Theta|}{\tau^\star(\alpha_0)} + \alpha_0\right)\right] \geq \frac{\mathbb{E}[\Theta^2]}{2M^2}\phi\left(\frac{M}{\tau^\star(\alpha_0)} + \alpha_0\right). \tag{57}$$

Also, it is easily seen that

$$|E_2| \leq \max_{x\in\mathbb{R}}\left\{x\phi(x - \alpha_0) - x\phi(-x - \alpha_0)\right\}, \tag{58}$$

where the right-hand side is positive and bounded away from 0 whenever α_0 is positive. Therefore, the right-hand side of the expression (55) and the expression (56) are both bounded away from 0 and ∞ for any fixed $\alpha_0 \geq \alpha_{\min}$. Combining these two cases, we have proved the advertised inequality (49).

B.3 Proof of Proposition B.3

From the proof of Bayati and Montanari (2011b, Proposition 1.3), we know that the function $\tau^2 \mapsto \mathsf{F}(\tau^2, \alpha\tau)$ is concave for any $\alpha > 0$ and Θ not equal to 0. Therefore, we obtain

$$\left|\frac{\tau_{t+1}^2 - \tau_t^{\star 2}}{\tau_t^2 - \tau_t^{\star 2}}\right| \leq \frac{\tau_t^{\star 2} - \sigma^2}{\tau_t^{\star 2}} \leq 1 - \frac{\sigma^2}{\tau_{\max}^{\star\ 2}} =: \eta,$$

as $\tau_{t+1}^2 = \mathsf{F}(\tau^2, \alpha\tau)$ and $\sigma^2 = \mathsf{F}(0, 0)$. Consequently, it leads to

$$\left|\tau_{t+1}^2 - \tau_{t+1}^{\star\ 2}\right| \leq \eta\left|\tau_t^2 - \tau_t^{\star 2}\right| + \left|\tau_t^{\star 2} - \tau_{t+1}^{\star\ 2}\right|.$$

Without loss of generality, assume $\{\lambda_t\}$ decays with t. Invoking the above relation recursively, we obtain

$$\frac{\left|\tau_{t+1}^2 - \tau_{t+1}^{\star\ 2}\right|}{\eta^{t+1}} \leq \frac{\left|\tau_1^2 - \tau_1^{\star 2}\right|}{\eta} + \sum_{s=1}^{t} \frac{\left|\tau_s^{\star 2} - \tau_{s+1}^{\star\ 2}\right|}{\eta^s}$$

$$
\begin{aligned}
(\text{since } \tau^\star(\lambda) \text{ is } L_\tau\text{-Lipschitz }) &\le \frac{\left|\tau_1^2-\tau_1^{\star 2}\right|}{\eta}+2\tau_{\max}^\star L_\tau \sum_{s=1}^{[t/2]}\frac{|\lambda_s-\lambda_{s+1}|}{\eta^s}+2\tau_{\max}^\star L_\tau \sum_{s=1+[t/2]}^{t}\frac{|\lambda_s-\lambda_{s+1}|}{\eta^s} \\
&\le \frac{\left|\tau_1^2-\tau_1^{\star 2}\right|}{\eta}+2\tau_{\max}^\star L_\tau \frac{1}{\eta^{[t/2]}}\sum_{s=1}^{[t/2]}[\lambda_s-\lambda_{s+1}]+2\tau_{\max}^\star L_\tau \frac{[\lambda_{[t/2]}-\lambda_{[t/2]+1}]}{1-\eta}\eta^{-t} \\
&\le \frac{\left|\tau_1^2-\tau_1^{\star 2}\right|}{\eta}+2\tau_{\max}^\star L_\tau \lambda_1 \frac{1}{\eta^{[t/2]}}+2\tau_{\max}^\star L_\tau \frac{[\lambda_{[t/2]}-\lambda_{[t/2]+1}]}{1-\eta}\eta^{-t}.
\end{aligned}
$$

Re-arranging the above expression, we are left with

$$
\left|\tau_t^2-\tau_t^{\star 2}\right| \le \frac{\left|\tau_1^2-\tau_1^{\star 2}\right|}{\eta}\eta^t+2\tau_{\max}^\star L_\tau \lambda_1 \eta^{[t/2]}+\frac{2\tau_{\max}^\star L_\tau}{1-\eta}[\lambda_{[t/2]}-\lambda_{[t/2]+1}] \to 0, \quad t\to\infty.
$$

This completes the proof of Proposition B.3.

C Proofs about the AMP updates

The goal of this section is to prove Theorem 3. In Section C.1, we state a key lemma (cf. Lemma C.1), which characterizes the convergence of the AMP updates (as $t\to\infty$) to the minimum ℓ_1-norm interpolator $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$. The proof of Theorem 3 is then built upon this lemma. Section C.2 is then devoted to the main proof of Lemma C.1 with auxiliary lemmas established in Section C.3 and Section C.4.

The main structure of the proof is similar to that of Bayati and Montanari (2011b); in the following text, we often refer to the paper as BM for simplification. The major difference between the min ℓ_1 scenario and a fixed λ scenario (considered in Bayati and Montanari (2011b) and other references) lies in the lack of restricted strong convexity around the solution point, which prevents us from translating the closeness in the loss function values to the proximity of the minimizers. We shall overcome this challenge by carefully investigating the AMP updates for decaying choices of the regularization parameter. Throughout this section, we make use of the properties for the state evolution parameters repeatedly (we refer the readers to Section B for more details).

C.1 Proof of Theorem 3

The key enabler for obtaining Theorem 3 from Proposition 2 is the following result, which connects the minimum ℓ_1-norm interpolator with the AMP iterations in expression (21) (with the parameters selected according to (27)).

Lemma C.1. *The sequence produced by the AMP updates satisfies*

$$
\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\|\boldsymbol{\theta}^t-\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2^2 \overset{\text{a.s.}}{=} 0. \tag{59}
$$

Let us first provide the proof of Theorem 3 and defer the proof of Lemma C.1 to Section C.2. To begin with, given any $t\ge 0$, in view of the pseudo-Lipschitz property of ψ (cf. (6)), one has

$$
\begin{aligned}
\left|\frac{1}{p}\sum_{i=1}^p \psi(\widehat{\theta}_i^{\mathsf{lnt}},\theta_i^\star)-\frac{1}{p}\sum_{i=1}^p \psi(\theta_i^{t+1},\theta_i^\star)\right| &\le \frac{L}{p}\sum_{i=1}^p |\theta_i^{t+1}-\widehat{\theta}_i^{\mathsf{lnt}}|\left(1+\sqrt{(\widehat{\theta}_i^{\mathsf{lnt}})^2+(\theta_i^\star)^2}+\sqrt{(\theta_i^{t+1})^2+(\theta_i^\star)^2}\right) \\
&\le \frac{L}{p}\|\boldsymbol{\theta}^{t+1}-\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2\sqrt{\sum_{i=1}^p\left(1+|\widehat{\theta}_i^{\mathsf{lnt}}|+|\theta_i^{t+1}|+2|\theta_i^\star|\right)^2} \\
&\le L\frac{\|\boldsymbol{\theta}^{t+1}-\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2}{\sqrt{p}}\cdot\sqrt{4+\frac{16\|\boldsymbol{\theta}^\star\|_2^2}{p}+\frac{4\|\boldsymbol{\theta}^{t+1}\|_2^2}{p}+\frac{4\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2^2}{p}}.
\end{aligned}
$$

Regarding the right-hand side of the above relation, Lemma C.1 guarantees that $\lim_{p\to\infty} \|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2/\sqrt{p} \to 0$ almost surely as $t \to \infty$; our assumptions about $\boldsymbol{\theta}^\star$ ensure that $\|\boldsymbol{\theta}^\star\|_2^2/p$ is bounded. By virtue of Lemma C.3, the other two terms involving state evolution $\boldsymbol{\theta}^t$ and the ℓ_1-minimization solution $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ are also bounded. Putting these together, one can readily conclude that

$$\begin{aligned}
\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \psi(\widehat{\theta}_i^{\mathsf{lnt}}, \theta_i^\star) &= \lim_{t\to\infty}\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \psi(\theta_i^t, \theta_i^\star) \quad \text{(Lemma C.1, Lemma C.3)} \\
&= \lim_{t\to\infty} \mathbb{E}\left\{\psi(\eta(\Theta + \tau_t Z; \zeta_t), \Theta)\right\} \quad \text{(Proposition 2)} \\
&= \mathbb{E}\left\{\psi(\eta(\Theta + \tau^\star Z; \alpha^\star\tau^\star), \Theta)\right\}.
\end{aligned}$$

Here, the last step makes use of Proposition B.3 and Proposition B.2 which demonstrates that $\tau_t \to \tau^\star$ and $\alpha_t^\star \to \alpha^\star$ as $t \to \infty$. The proof of Theorem 3 is thus complete.

C.2 Proof of Lemma C.1

This section is devoted to the proof of Lemma C.1. To this end, we first introduce a key result in Lemma C.2 which characterizes the conditions under which, the ℓ_2-norm of the perturbation $\|\boldsymbol{r}\|_2$ can be controlled whenever the difference between $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta} + \boldsymbol{r})$ and $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta})$ (cf. (60)) is small. The proof of this lemma can be found in Section C.3.1. With Lemma C.2 in place, proving Lemma C.1 boils down to verifying each required condition. To accomplish this goal, we make use of a series of results in Lemmas C.3-C.6, followed by the complete proof of Lemma C.1. The proofs of these auxiliary results are deferred to Section C.3.

We define the Lasso problem associated with the regularization parameter λ_t as follows

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}(\lambda_t; \boldsymbol{X}, \boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{C}_{\lambda_t}(\boldsymbol{\theta}), \quad \text{where} \quad \mathcal{C}_{\lambda_t}(\boldsymbol{\theta}) \coloneqq \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda_t\|\boldsymbol{\theta}\|_1. \tag{60}$$

As we shall make clear momentarily, each iterate of our AMP updates aims to take a step closer to the minimizer of $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta})$. More connections with these Lasso estimators shall be pointed out in the sequel.

Lemma C.2. *There exists a function $\xi(\varepsilon, c_1, \ldots, c_6)$ such that, if $\boldsymbol{\theta}, \boldsymbol{r} \in \mathbb{R}^p$ satisfy the following conditions:*

1. $\|\boldsymbol{r}\|_2 \le c_1\sqrt{p}$;
2. $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta} + \boldsymbol{r}) \le \mathcal{C}_{\lambda_t}(\boldsymbol{\theta}) + c_2\lambda_t p\varepsilon$;
3. *There exists* $\mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}) \in \partial\mathcal{C}_{\lambda_t}(\boldsymbol{\theta})$, *such that* $\|\mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta})\|_2 \le \sqrt{p}\lambda_t\varepsilon$;
4. *Let* $\boldsymbol{s} = (1/\lambda_t)[\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + sg(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta})] \in \partial\|\boldsymbol{\theta}\|_1$, *and* $S(c_3) = \{i \in [p] : |s_i| \ge 1 - c_3\}$. *Then for any* $S' \subseteq [p]$, $|S'| \le c_4 p$, *we have* $\sigma_{\min}(\boldsymbol{X}_{S(c_3)\cup S'}) \ge c_5$;
5. $\sigma_{\max}(\boldsymbol{X}) \le c_6$,

then $\|\boldsymbol{r}\|_2 \le \sqrt{p}\xi(\varepsilon, c_1, \ldots, c_6)$, *and for any* $c_1, \ldots, c_6 > 0$, *one has* $\xi(\varepsilon, c_1, \ldots, c_6) \to 0$ *as* $\varepsilon \to 0$.

A few remarks are in order. First, we write $\xi(\varepsilon, c_1, \ldots, c_6)$ to emphasize that the function does not depend on $\boldsymbol{X}$, $\mathcal{C}_{\lambda_t}$ or the constructions of $\boldsymbol{\theta}, \boldsymbol{r}$. Secondly, Lemma C.2 is a generalization of BM-Lemma 3.1. Since BM is concerned with the Lasso estimator with a single positive λ, the corresponding lemma only requires $\mathcal{C}_\lambda(\boldsymbol{\theta} + \boldsymbol{r}) \le \mathcal{C}_\lambda(\boldsymbol{\theta})$, and is employed for the Lasso loss function with $\boldsymbol{\theta} + \boldsymbol{r}$ being the Lasso solution. In this case, $\mathcal{C}_\lambda(\boldsymbol{\theta} + \boldsymbol{r}) \le \mathcal{C}_\lambda(\boldsymbol{\theta})$ holds true for every $\boldsymbol{\theta}$ by definition. In our setting, however, the minimum ℓ_1-norm solution $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ is not the minimizer of $\mathcal{C}_{\lambda_t}$ — recognizing the fact that we aim to apply Lemma C.2 with $\boldsymbol{\theta} + \boldsymbol{r} = \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$. BM-Lemma 3.1 therefore does not apply directly. Hence, it requires us to generalize their lemma in a way suitable to our setting.

Proof of Lemma C.1. Now suppose that one can find constants $(c_1, \ldots, c_6)$ and a sequence $\{\varepsilon_t\}$ satisfying $\lim_{t\to\infty} \varepsilon_t = 0$, such that the five conditions in Lemma C.2 hold almost surely with the choice of $\varepsilon = \varepsilon_t$, $\boldsymbol{\theta} = \boldsymbol{\theta}^t$ and $\boldsymbol{\theta} + \boldsymbol{r} = \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$. As a result, we know that for each t, it holds that

$$\|\boldsymbol{\theta}^t - \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2 \le \sqrt{p}\xi(\varepsilon_t, c_1, \ldots, c_6)$$

almost surely. Further taking $t \to \infty$ yields

$$\lim_{t\to\infty} \frac{\|\boldsymbol{\theta}^t - \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2}{\sqrt{p}} \le \lim_{t\to\infty} \xi(\varepsilon_t, c_1, \ldots, c_6) = 0,$$

where the last equality follows from $\lim_{t\to\infty} \varepsilon_t = 0$ and the conclusion in Lemma C.2 that $\xi(\varepsilon, c_1, \ldots, c_6) \to 0$ as $\varepsilon \to 0$. Thus we complete the proof of Lemma C.1. It remains to construct these $(c_1, \ldots, c_6)$ and a converging sequence of ε_t, such that the five conditions in Lemma C.2 hold almost surely as $p \to \infty$.

To begin with, we make the observation that the last condition of Lemma C.2 follows directly from the classical random matrix theory where $\sigma_{\max}(\boldsymbol{X}) \to 1 + \frac{1}{\sqrt{\delta}}$ (see, e.g., Bai and Silverstein (2010)). Thus it suffices to verify the other four conditions.

- *Condition 1 of Lemma C.2.* We introduce the following lemma, which develops upper bounds on the ℓ_2-norm of both the ℓ_1-interpolation solution and the AMP iterations. The proof of this result is deferred to Section C.3.2.

  Lemma C.3. *There exists a constant C that only depends on Θ, σ and δ, such that, almost surely*

  $$\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{p}\|\boldsymbol{\theta}^t\|_2^2 < C; \tag{61a}$$

  $$\lim_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2^2 < C. \tag{61b}$$

  By virtue of this lemma, there exists a constant $c_1 > 0$ such that $\|\boldsymbol{r}\|_2 = \|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}} - \boldsymbol{\theta}^t\|_2 \le c_1\sqrt{p}$, which validates Condition 1 of Lemma C.2.

- *Condition 3 of Lemma C.2.* Similar to the constructions in BM (pg. 25), let us denote

  $$s_i^t := \begin{cases} \operatorname{sign}(\theta_i^t), & \text{if } \theta_i^t \neq 0; \\ \frac{1}{\zeta_{t-1}}\left\{[\boldsymbol{X}^\top \boldsymbol{z}^{t-1}]_i + \theta_i^{t-1}\right\}, & \text{otherwise,} \end{cases} \quad \text{or} \quad \boldsymbol{s}^t := \frac{1}{\zeta_{t-1}}\left(\boldsymbol{\theta}^{t-1} + \boldsymbol{X}^\top \boldsymbol{z}^{t-1} - \boldsymbol{\theta}^t\right). \tag{62}$$

  In view of the AMP updates (21a), we can verify the equivalence of these two expressions above, and that vector

  $$\operatorname{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t) := \lambda_t \boldsymbol{s}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^t) \tag{63}$$

  is a valid sub-gradient of $\mathcal{C}_{\lambda_t}$ at $\boldsymbol{\theta}^t$. With these notation in place, we introduce the following lemma which controls the ℓ_2-norm of $\operatorname{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t)$. The proof of this result can be found in Section C.3.3.

  Lemma C.4. *Under our choices of $\{\lambda_t\}$ and $\{\zeta_t\}$, the sub-gradient of $\mathcal{C}_{\lambda_t}$ at point $\boldsymbol{\theta}^t$ defined in (63) satisfy*

  $$\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{\sqrt{p}\lambda_t}\|\operatorname{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t)\|_2 \overset{\text{a.s.}}{=} 0. \tag{64}$$

  In other words, there exists a sequence of $\{\varepsilon_t\}$ approaching zero such that: for each t, it holds almost surely that $\|\operatorname{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t)\|_2 \le \sqrt{p}\lambda_t\varepsilon_t$.

- *Condition 4 in Lemma C.2.* As defined above, the vector $\boldsymbol{s}^t$ in the expression (62) satisfies

  $$\boldsymbol{s}^t = \lambda_t^{-1}\left[\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^t) + \operatorname{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t)\right] \in \partial\|\boldsymbol{\theta}^t\|_1,$$

  where $\operatorname{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t) \in \partial\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}^t)$ (cf. (63)) is a valid sub-gradient of $\mathcal{C}_{\lambda_t}$. It turns out that the approximate support set of $\boldsymbol{\theta}^t$ does not vary too much across iterations of the AMP algorithm. This is a high-level reason why $\sigma_{\min}(\boldsymbol{X}_{S(c_3)\cup S'})$ can be bounded away from zero, despite the fact that the loss function is not strongly convex. This observation is rigorized in the lemma below.

Lemma C.5. *Given every $\gamma \in (0,1)$ and $t \geq 1$, define the set*

$$S_t(\gamma) = \left\{ i \in [p] : |s_i^t| \geq 1 - \gamma \right\}, \tag{65}$$

with $\boldsymbol{s}^t$ defined in (62). Then for any $\xi > 0$ there exists $t_\star = t_\star(\xi, \gamma) < \infty$ such that, for all $t_2 \geq t_1 \geq t_\star$, one has

$$|S_{t_2}(\gamma)/S_{t_1}(\gamma)| < p\xi$$

almost surely as $p \to \infty$.

The proof of Lemma C.5 is provided in Section C.4. Recall that BM-Lemma C.5 establishes similar results for the AMP iterates corresponding to a fixed λ. Here, we aim to approximate the Lasso solution for different parameter λ_t at each step of the AMP iteration. Therefore it requires us to consider auxiliary state-evolution formulas for each t separately in order to establish Lemma C.5.

Based on Lemma C.5, one can derive the following result concerning the constrained singular value of $\boldsymbol{X}$.

Lemma C.6. *There exists constraints $\gamma_1 \in (0,1)$, $\gamma_2, \gamma_3 > 0$ and $t_{\min} < \infty$ such that for any $t \geq t_{\min}$,*

$$\min_{S'} \left\{ \sigma_{\min}(\boldsymbol{X}_{S_t(\gamma_1) \cup S'}) : S' \subseteq [p], |S'| \leq \gamma_2 p \right\} \geq \gamma_3, \tag{66}$$

almost surely as $p \to \infty$.

Based on the conclusion of Lemma C.3, the proof of Lemma C.6 follows verbatim as of BM-Proposition 3.6, and is thus omitted here. Note that apart from Lemma C.3, the proof requires BM-Lemma 3.4, which holds for AMP iterations with general choices of $\{\zeta_t\}$, which can be directly adapted to accommodate our setting.

As a direct consequence of Lemma C.6, Condition 4 in Lemma C.2 follows immediately with the choice of $\boldsymbol{s} = \boldsymbol{s}^t$.

- *Condition 2 in Lemma C.2.* To verify this condition, we only need to find a vanishing sequence $\{\varepsilon_t\}$ such that $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}^t) - \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}) \geq -c_2 \lambda_t p \varepsilon_t$. Towards this end, we shall control $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}^t) - \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}})$ as follows. First, recalling that $\widehat{\boldsymbol{\theta}}_t$ is the minimizer of $\mathcal{C}_{\lambda_t}(\cdot)$ (cf. (60)). Hence,

$$\begin{aligned} \frac{\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}^t) - \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}})}{p} \geq \frac{\mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}_t) - \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}})}{p} &= \frac{1}{2p}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_t\|_2^2 + \frac{\lambda_t}{p}\left[\|\widehat{\boldsymbol{\theta}}_t\|_1 - \|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1\right] \\ &\geq \frac{\lambda_t}{p}\left[\|\widehat{\boldsymbol{\theta}}_t\|_1 - \|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1\right]. \end{aligned} \tag{67}$$

In the following, we aim to show that $\|\widehat{\boldsymbol{\theta}}_t\|_1 - \|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1 \geq -c_2\varepsilon_t$ for some vanishing sequence $\{\varepsilon_t\}$. From the definition of $\mathcal{C}_{\lambda_t}$ in the expression (60), for every $\boldsymbol{\theta}$ we can express

$$\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}) = \mathcal{C}_{\lambda_{t+1}}(\boldsymbol{\theta}) + (\lambda_t - \lambda_{t+1})\|\boldsymbol{\theta}\|_1;$$

both of the right-hand side terms $\mathcal{C}_{\lambda_{t+1}}(\boldsymbol{\theta})$ and $(\lambda_t - \lambda_{t+1})\|\boldsymbol{\theta}\|_1$ are convex and non-negative functions of $\boldsymbol{\theta}$, and are minimized at $\widehat{\boldsymbol{\theta}}_{t+1}$ and $\mathbf{0}_p$, respectively. For any $\boldsymbol{\theta} \in \mathbb{R}^p$ with $\|\boldsymbol{\theta}\|_1 > \|\widehat{\boldsymbol{\theta}}_{t+1}\|_1 \geq 0$, it is easily seen that $\mathcal{C}_{\lambda_{t+1}}(\boldsymbol{\theta}) \geq \mathcal{C}_{\lambda_{t+1}}(\widehat{\boldsymbol{\theta}}_{t+1})$ and $(\lambda_t - \lambda_{t+1})\|\boldsymbol{\theta}\|_1 > (\lambda_t - \lambda_{t+1})\|\widehat{\boldsymbol{\theta}}_{t+1}\|_1$; it follows that $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}) > \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}_{t+1}) \geq \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}_t)$. Thus we reach the conclusion that $\|\widehat{\boldsymbol{\theta}}_t\|_1 \leq \|\widehat{\boldsymbol{\theta}}_{t+1}\|_1$. Additionally, by virtue of Lemma C.3, we obtain that

$$\lim_{p\to\infty} \frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1}{p} \leq \lim_{p\to\infty} \frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2}{\sqrt{p}} \leq \sqrt{C}$$

is bounded almost surely. As a summery, the sequence $\|\widehat{\boldsymbol{\theta}}_t\|_1/p$ enjoys the following two properties

– $\|\widehat{\boldsymbol{\theta}}_t\|_1/p \le \|\widehat{\boldsymbol{\theta}}_{t+1}\|_1/p \le \cdots \le \|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1/p$, and almost surely as $p \to \infty$, $\lim_{t\to\infty} \widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$;

– $\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1/p$ is bounded almost surely.

In view of the monotone convergence theorem, we reach the conclusion that

$$\lim_{t\to\infty} \frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1 - \|\widehat{\boldsymbol{\theta}}_t\|_1}{p} = 0 \quad \Longrightarrow \quad \frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1 - \|\widehat{\boldsymbol{\theta}}_t\|_1}{p} \le \varepsilon_t,$$

for a vanishing sequence $\{\varepsilon_t\}$. Combining with expression (67), we arrive at $\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}^t) - \mathcal{C}_{\lambda_t}(\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}) \ge -c_2\lambda_t p\varepsilon_t$. Thus we verify the condition 2 in Lemma C.2.

Taking the above results collectively, we complete the proof of Lemma C.1.

C.3 Proof of supporting lemmas to Lemma C.1

C.3.1 Proof of Lemma C.2

As experienced readers might have already noticed, the statement of Lemma C.2 is very similar to that of BM-Lemma 3.1; the only difference lies in the Condition 2 and Condition 3. A closer inspection at the proof of BM-Lemma 3.1 reveals that these two conditions are only used to establish BM-(3.4) and BM-(3.5). Therefore, as long as we can show BM-(3.4) and BM-(3.5) for our setting, the proof of our lemma will be completed, with the rest part following verbatim from the proof of BM-Lemma 3.1.

Let us first adapt BM-(3.4) and BM-(3.5) to our setting. Throughout, we shall use $\xi(\varepsilon)$ to denote a function of constants $c_1, \ldots, c_6 > 0$ and of ε such that $\xi(\varepsilon) \to 0$ as $\varepsilon \to 0$. Additionally, we shall use $S = \mathrm{supp}(\boldsymbol{\theta}) \subseteq [p]$. Formally, it suffices for us to show that, one can find such $\xi(\varepsilon)$ where

$$\frac{\|\boldsymbol{r}_{\bar{S}}\|_1 - \langle \boldsymbol{s}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle}{p} \le \xi(\varepsilon); \tag{68a}$$

$$\|\boldsymbol{X}\boldsymbol{r}\|_2^2 \le p\xi(\varepsilon). \tag{68b}$$

Given these two inequalities, the proof of Lemma C.2 follows directly from BM-Lemma 3.1. For the sake of brevity, we only establish the aforementioned two inequalities, and refer readers to Bayati and Montanari (2011b) for the rest of the proof.

Verifying the expressions (68a) and (68b). With $\boldsymbol{s}$ defined in condition 4, we obtain

$$\begin{aligned}
c_2\varepsilon \ge & \frac{\mathcal{C}_{\lambda_t}(\boldsymbol{\theta}+\boldsymbol{r}) - \mathcal{C}_{\lambda_t}(\boldsymbol{\theta})}{p\lambda_t} \quad \text{(Condition 2)} \\
= & \frac{\|\boldsymbol{\theta}_S + \boldsymbol{r}_S\|_1 - \|\boldsymbol{\theta}_S\|_1}{p} + \frac{\|\boldsymbol{r}_{\bar{S}}\|_1}{p} + \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{r}\|_2^2 - \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2p\lambda_t} \\
\overset{\text{(i)}}{=} & \frac{\|\boldsymbol{\theta}_S + \boldsymbol{r}_S\|_1 - \|\boldsymbol{\theta}_S\|_1 - \langle \mathrm{sign}(\boldsymbol{\theta}_S), \boldsymbol{r}_S\rangle}{p} + \frac{\|\boldsymbol{r}_{\bar{S}}\| - \langle \boldsymbol{s}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle}{p} + \frac{\lambda_t\langle \boldsymbol{s}, \boldsymbol{r}\rangle - \langle \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{r}\rangle + \frac{1}{2}\|\boldsymbol{X}\boldsymbol{r}\|_2^2}{p\lambda_t} \\
\overset{\text{(ii)}}{=} & \frac{\|\boldsymbol{\theta}_S + \boldsymbol{r}_S\|_1 - \|\boldsymbol{\theta}_S\|_1 - \langle \mathrm{sign}(\boldsymbol{\theta}_S), \boldsymbol{r}_S\rangle}{p} + \frac{\|\boldsymbol{r}_{\bar{S}}\| - \langle \boldsymbol{s}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle}{p} + \frac{\langle \mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}), \boldsymbol{r}\rangle}{p\lambda_t} + \frac{\|\boldsymbol{X}\boldsymbol{r}\|_2^2}{2p\lambda_t}. \quad \text{(Condition 4).}
\end{aligned}$$

Here, (i) follows from the fact that $\boldsymbol{s} \in \partial\|\boldsymbol{\theta}\|_1$, and thus $\mathrm{sign}(\boldsymbol{\theta}_S) = \boldsymbol{s}_S$. The equality (ii) comes from the definition (63).

Invoking the Cauchy-Schwarz inequality and Condition 3 yields $|\langle \mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}), \boldsymbol{r}\rangle/(p\lambda_t)| \le c_1\varepsilon$. Substitution into the above inequality with a little algebra leads to

$$\frac{\|\boldsymbol{\theta}_S + \boldsymbol{r}_S\|_1 - \|\boldsymbol{\theta}_S\|_1 - \langle \mathrm{sign}(\boldsymbol{\theta}_S), \boldsymbol{r}_S\rangle}{p} + \frac{\|\boldsymbol{r}_{\bar{S}}\| - \langle \boldsymbol{s}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle}{p} + \frac{\|\boldsymbol{X}\boldsymbol{r}\|_2^2}{2p\lambda_t} \le (c_1 + c_2)\varepsilon. \tag{69}$$

It can be easily seen that these three terms on the left-hand side above are all non-negative. Therefore, the equalities (68a) and (68b) follow directly.

C.3.2 Proof of Lemma C.3

The proof of this lemma is adapted from BM-Lemma 3.2 with modifications tailored to the minimum ℓ_1-norm solution. As mentioned previously, Proposition 2 holds for general choices of $\{\zeta_t\}$, and is therefore directly applicable to our setting. Hence, the AMP iterates satisfy

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\|\boldsymbol{\theta}^t\|_2^2 = \lim_{t\to\infty}\mathbb{E}\left[\eta^2(\Theta+\tau_t Z;\zeta_t)\right] \le \mathbb{E}[\Theta^2]+\tau^{\star 2}.$$

In other words, asymptotically $\frac{1}{p}\|\boldsymbol{\theta}^t\|_2^2$ is bounded by some constant that depends only on $\mathbb{E}[\Theta^2]$ and $\tau^\star$.

Consequently, to prove Lemma C.3, it suffices to establish the relation (61b). Let us first decompose the minimum ℓ_1-norm interpolator $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ into the projection onto the row space of $\boldsymbol{X}$, and the residual. Formally, we write

$$\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}} = \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{\perp} + \widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||},$$

where $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$. It is straightforward to verify that $\boldsymbol{X}\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||} = \boldsymbol{X}\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}} = \boldsymbol{y}$. We can view $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{\perp}$ as the projection of $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ onto the orthogonal space of $\mathrm{row}(\boldsymbol{X})$, which is a uniformly random $(p-n)$-dimensional subspace of $\mathbb{R}^p$. By Kashin Theorem (BM-Lemma F.1), there exists a universal constant $c_1>0$ depending on δ such that $\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{\perp}\|_2^2 \le c_1\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{\perp}\|_1^2/p$, almost surely as $p\to\infty$. In addition, regarding the limiting value for the eigenvalues of Wishart matrices, it is known that there exists a constant c_2 depending only on δ such that $\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||}\|_2^2 \le c_2\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||}\|_2^2$ almost surely as $p\to\infty$ (see BM-Lemma F.2). Therefore, we arrive at

$$\begin{aligned}
\lim_{p\to\infty}\frac{1}{p}\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2^2 &= \lim_{p\to\infty}\left[\frac{1}{p}\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{\perp}\|_2^2+\frac{1}{p}\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||}\|_2^2\right]\\
&\le \lim_{p\to\infty}\left[c_1\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{\perp}\|_1^2}{p^2}+\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||}\|_2^2}{p}\right]\\
(\text{by Cauchy-Schwarz}) &\le \lim_{p\to\infty}\left[2c_1\left(\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1^2}{p^2}+\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||}\|_2^2}{p}\right)+\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||}\|_2^2}{p}\right]\\
&\le \lim_{p\to\infty}\left[2c_1\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1^2}{p^2}+(2c_1+1)c_2\frac{\|\boldsymbol{y}\|_2^2}{p}\right],
\end{aligned}\tag{70}$$

where the last step uses $\boldsymbol{X}\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}_{||} = \boldsymbol{X}\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}} = \boldsymbol{y}$.

As a consequence, it suffices to upper bound the quantities $\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1/p$ and $\|\boldsymbol{y}\|_2^2/p$. First, by our model assumption, $y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\|\boldsymbol{\theta}^\star\|_2^2/n+\sigma^2)$; it immediately follows that $\lim_{p\to\infty}\|\boldsymbol{y}\|_2^2/p$ is bounded. In addition, we note that $\boldsymbol{a} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ satisfies the condition $\boldsymbol{X}\boldsymbol{a}=\boldsymbol{y}$. Since $\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}$ has the minimum ℓ_1-norm over all linear interpolators, we can guarantee that

$$\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1 \le \|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}\|_1.$$

Finally, by virtue of BM-Lemma F.2, we obtain $\sigma_{\max}((\boldsymbol{X}\boldsymbol{X}^\top)^{-1}) \le c_3$ for some $c_3>0$ depending on δ, almost surely as $p\to\infty$. Collecting these components together yields

$$\left(\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1}{p}\right)^2 \le \frac{1}{p}\|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}\|_2^2 \le \frac{\boldsymbol{y}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}}{p} \le 2\frac{\|\boldsymbol{\theta}^\star\|_2^2}{p}+2c_3\frac{\|\boldsymbol{z}\|_2^2}{p}.$$

Therefore, $\lim_{p\to\infty}(\frac{\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_1}{p})^2 \le 2\mathbb{E}[\Theta^2]+2c_3\sigma^2$ holds almost surely. Combining this with the expression (70), we see that $\lim_{p\to\infty}\frac{1}{p}\|\widehat{\boldsymbol{\theta}}^{\mathsf{lnt}}\|_2^2$ is upper bounded by a universal constant that depends only on δ.

Putting these two parts together finishes the proof of Lemma C.3.

C.3.3 Proof of Lemma C.4

To prove Lemma C.4, we aim to upper bound the target quantity $\|\mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t)\|_2$ by three terms, and look at each term separately. For ease of exposition, let us define

$$\omega_t = \frac{1}{\delta}\left\langle \eta'(\boldsymbol{X}^\top \boldsymbol{z}^{t-1} + \boldsymbol{\theta}^{t-1}; \zeta_{t-1})\right\rangle.$$

The AMP iterate (21b) then translates into $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^t = \boldsymbol{z}^t - \omega_t \boldsymbol{z}^{t-1}$. With this piece of notation in mind, we can rewrite the sub-gradient (cf. (63)) as

$$\begin{aligned}
\mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t) &= \lambda_t \boldsymbol{s}^t - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) - (1-\omega_t)\boldsymbol{X}^\top \boldsymbol{z}^{t-1} \\
&= \frac{\lambda_t}{\zeta_{t-1}}\left[\zeta_{t-1}\boldsymbol{s}^t - \boldsymbol{X}^\top \boldsymbol{z}^{t-1}\right] - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) + \frac{[\lambda_t - \zeta_{t-1}(1-\omega_t)]}{\zeta_{t-1}}\boldsymbol{X}^\top \boldsymbol{z}^{t-1} \\
&= \frac{\lambda_t}{\zeta_{t-1}}(\boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^t) - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) + \frac{[\lambda_t - \zeta_{t-1}(1-\omega_t)]}{\zeta_{t-1}}\boldsymbol{X}^\top \boldsymbol{z}^{t-1}.
\end{aligned}$$

Applying the triangle inequality leads to

$$\frac{1}{\sqrt{p}\lambda_t}\|\mathrm{sg}(\mathcal{C}_{\lambda_t}, \boldsymbol{\theta}^t)\|_2 \le \frac{1}{\zeta_{t-1}}\frac{\|\boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^t\|_2}{\sqrt{p}} + \sigma_{\max}(\boldsymbol{X})\frac{\|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|_2}{\sqrt{p}\lambda_t} + \sigma_{\max}(\boldsymbol{X})\frac{\|\boldsymbol{z}^{t-1}\|_2}{\sqrt{p}}\frac{[\lambda_t - \zeta_{t-1}(1-\omega_t)]}{\zeta_{t-1}\lambda_t}. \tag{71}$$

The proof of Lemma C.4 then boils down to analyzing the three terms on the right-hand side of (71). We first invoke the following lemma describing the convergence of the AMP updates, which will be proved in Section C.4.3.

Lemma C.7. *The AMP iterates obey the following convergence guarantees*

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p\lambda_t^2}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2^2 = 0, \qquad \lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p\lambda_t^2}\|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|_2^2 = 0.$$

Note that, in the limit, Proposition B.2 ensures that $\zeta_t \to \zeta^\star$, and is thus bounded away from 0. Combining Lemma C.7 and the choice of λ_t ensures $\frac{1}{\zeta_{t-1}}\frac{\|\boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^t\|_2}{\sqrt{p}}$ converges to zero as $t \to \infty$. In addition, since $\sigma_{\max}(\boldsymbol{X})$ is almost surely bounded in the limit, the second term on the right-hand side of the expression (71) also converges to 0.

It remains to consider the third term on the right-hand side of the expression (71). To this end, one can first characterize the large-system limit of $\|\boldsymbol{z}^t\|_2$ as follows

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{\|\boldsymbol{z}^t\|_2}{\sqrt{p}} = \lim_{t\to\infty}\tau_t = \tau^\star, \quad \text{almost surely}, \tag{72}$$

as shown in BM-Lemma 4.1. In addition, the result in BM-Lemma F.3 demonstrates that $\forall t \ge 1$,

$$\lim_{p\to\infty}[1-\omega_t] \overset{\text{a.s.}}{=} 1 - \frac{1}{\delta}\mathbb{E}\left[\eta'(\Theta + \tau_t Z; \zeta_t)\right].$$

By construction, we know $\lambda_t = \tau_t^\star \alpha_t^\star\left(1 - \frac{1}{\delta}\mathbb{E}\left[\eta'(\Theta + \tau_t^\star Z; \tau_t^\star \alpha_t^\star)\right]\right)$. Hence,

$$\begin{aligned}
\left|\frac{1}{\alpha_t^\star \tau_t^\star} - \frac{1-\omega_t}{\lambda_t}\right| &\overset{\text{a.s.}}{\to} \frac{1}{\lambda_t}\left|\frac{\lambda_t}{\alpha_t^\star\tau_t^\star} - \left(1 - \frac{1}{\delta}\mathbb{E}\left[\eta'(\Theta + \tau_t Z; \tau_t\alpha_t^\star)\right]\right)\right| \\
&= \frac{1}{\delta\lambda_t}\left|\mathbb{E}\left[\eta'(\Theta + \tau_t^\star Z; \alpha_t^\star\tau_t^\star)\right] - \mathbb{E}\left[\eta'(\Theta + \tau_t Z; \tau_t\alpha_t^\star)\right]\right| \\
&\le \frac{1}{\delta\lambda_t}\mathbb{E}\left[\left|\Phi\left(\alpha_t^\star - \frac{\Theta}{\tau_t^\star}\right) - \Phi\left(\alpha_t^\star - \frac{\Theta}{\tau_t}\right)\right|\right] + \frac{1}{\delta\lambda_t}\mathbb{E}\left[\left|\Phi\left(\alpha_t^\star + \frac{\Theta}{\tau_t^\star}\right) - \Phi\left(\alpha_t^\star + \frac{\Theta}{\tau_t}\right)\right|\right] \\
&\le \frac{2\mathbb{E}[|\Theta|]}{\delta\tau_t^\star\tau_t}\frac{|\tau_t - \tau_t^\star|}{\lambda_t},
\end{aligned}$$

where the second inequality comes from the Lipschitz property of $\Phi(\cdot)$. If we take $t \to \infty$, then the limiting values of $\tau_t^\star \to \tau^\star$ and $\tau_t \to \tau^\star$ and Lemma C.9 immediately indicate that

$$\frac{2\mathbb{E}[|\Theta|]}{\delta \tau_t^\star \tau_t} \overset{\text{a.s.}}{\to} \frac{2\mathbb{E}[|\Theta|]}{\delta \tau^{\star 2}}; \quad \frac{|\tau_t - \tau_t^\star|}{\lambda_t} \overset{\text{a.s.}}{\to} 0.$$

Combining this with the model construction that $\mathbb{E}[|\Theta|] < +\infty$, we obtain

$$(1 - \omega_t)/\lambda_t \overset{\text{a.s.}}{\to} 1/(\alpha^\star \tau^\star),$$

and it follows that

$$\lim_{t\to\infty} \lim_{p\to\infty} \frac{[\lambda_t - \zeta_{t-1}(1-\omega_t)]}{\zeta_{t-1}\lambda_t} \overset{\text{a.s.}}{=} \lim_{t\to\infty} \left[\frac{1}{\zeta_{t-1}} - \frac{1}{\alpha^\star \tau^\star}\right] = 0.$$

Taken collectively with the expression (72), the third term of the inequality (71) converges to zero. This concludes the proof.

C.4 Proof of Lemma C.5 and Lemma C.7

The goal of this section is to prove Lemma C.5 and Lemma C.7, which requires us to characterize the changes between two AMP iterates $\boldsymbol{\theta}^s$ and $\boldsymbol{\theta}^t$ in different iterations $s \neq t$. Towards this end, in Section C.4.1, we define the covariance between AMP iterates, and state two auxiliary lemmas about the convergence rate of these covariances (cf. Lemma C.9 and Lemma C.10). In Section C.4.2 and Section C.4.3, we invoke these two lemmas to derive Lemma C.5 and Lemma C.7 respectively, which is then followed by the proofs of these two lemmas.

C.4.1 Auxiliary definitions and lemmas

By virtue of Proposition 2, the state evolution sequence $\{\tau_t^2\}_{t=0}^\infty$ can be viewed as the large n "variance" of the AMP recursions. We generalize this notion to consider the correlations between $\boldsymbol{\theta}^s$ and $\boldsymbol{\theta}^t$ when $s \neq t$. Formally, define a sequence of scalars $\{\mathsf{R}_{s,t}\}_{s,t\geq 0}$ recursively as in BM-(4.13):

$$\mathsf{R}_{s+1,t+1} \coloneqq \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta(\Theta + Z_s; \zeta_s) - \Theta]\,[\eta(\Theta + Z_t; \zeta_t) - \Theta]\right\}. \tag{73}$$

Here $(Z_s, Z_t) \in \mathbb{R}^2$ are jointly Gaussian, independent of Θ, with zero mean and

$$\mathbb{E}[Z_s^2] = \mathsf{R}_{s,s}, \quad \mathbb{E}[Z_t^2] = \mathsf{R}_{t,t}, \quad \mathbb{E}[Z_s Z_t] = \mathsf{R}_{s,t}. \tag{74}$$

The boundary conditions are given by $\mathsf{R}_{0,0} = \sigma^2 + \mathbb{E}[\Theta^2]/\delta$, and

$$\mathsf{R}_{0,t+1} = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta(\Theta + Z_t; \zeta_t) - \Theta]\,(-\Theta)\right\}, \tag{75}$$

where $Z_t \sim \mathcal{N}(0, \mathsf{R}_{t,t})$ is independent of Θ. Comparing these with the definition of state evolution formula in expressions (24a) and (24b), we can immediately see that $\mathsf{R}_{t,t} = \tau_t^2$ for $t \geq 0$. We record the following result from Bayati and Montanari (2011b).

Lemma C.8 (Theorem 4.2 in Bayati and Montanari (2011b))**.** *Under the setting of Proposition 2, given any pseudo-Lipschitz function $\psi : \mathbb{R}^3 \to \mathbb{R}$, it holds that*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^p \psi(\theta_i^s + (\boldsymbol{X}^\top \boldsymbol{z}^s)_i, \theta_i^t + (\boldsymbol{X}^\top \boldsymbol{z}^t)_i, \theta_i^\star) \overset{\text{a.s.}}{=} \mathbb{E}\left\{\psi(\Theta + Z_s, \Theta + Z_t, \theta_i^\star)\right\},$$

where (Z_s, Z_t) are jointly Gaussian, independent from Θ, mean-zero and satisfy (74).

Before embarking on the proofs of Lemma C.5 and Lemma C.7, let us first introduce two crucial lemmas concerning the convergence rate of $\mathsf{R}_{s,t}$ for both the cases of $s = t$ and $s \neq t$ (note that the $s = t$ case represents the convergence rate of τ_t^2). Their proofs can be found in Section C.4.5 and Section C.4.4, respectively.

Lemma C.9 (Convergence rate of the variance and 1-step covariance). *Define* $\mathsf{R}_{s,t}$ *as in then expression* (73). *We have*

$$\max\left\{|\mathsf{R}_{t,t}-\tau_t^{\star 2}|, |\mathsf{R}_{t+1,t+1}-\tau_t^{\star 2}|, |\mathsf{R}_{t,t}-2\mathsf{R}_{t,t+1}+\mathsf{R}_{t+1,t+1}|\right\} \le c_0\exp\{-c\Lambda_t\}+4L_\tau\tau^\star_{\max}l_t, \tag{76}$$

where $l_t \coloneqq \sum_{s=1}^t |\lambda_s-\lambda_{s+1}|\exp\{-c[\Lambda_t-\Lambda_s]\}$ *as in Assumption 1. Here* $\tau_t^\star$ *is the solution to the system of equations* (15), $c=(2\alpha^\star_{\max}\tau^\star_{\max})^{-1}$, L_τ *is defined in Proposition B.1, and* $c_0>0$ *is some constant that depends on* Θ, σ *and* δ.

With Lemma C.9 in place, for general t_1 and t_2, one can easily decompose R_{t_1,t_2} into multiple consecutive differences and obtain the respective convergence rates as follows.

Lemma C.10 (Convergence rate of general covariances). *For any* $t_0\ge 0$ *and* $t_1,t_2\ge t_0$, *with the same* c_0 *and* c *defined in Lemma C.9, the covariance* R_{t_1,t_2} *satisfies*

$$|\mathsf{R}_{t_1,t_1}-2\mathsf{R}_{t_1,t_2}+\mathsf{R}_{t_2,t_2}| \le 4\left[\sum_{i=t_1}^\infty \sqrt{c_0\exp\{-c\Lambda_i\}+4L_\tau\tau^\star_{\max}l_i}\right]^2.$$

Proposition B.3 ensures that $\mathsf{R}_{t_i,t_i}=\tau_{t_i}^2\to\tau^{\star 2}$ as $t_0\to\infty$. Under Assumption 1 (so that $\sum_{t=1}^\infty\sqrt{l_t}<+\infty$ and $\sum_{t=1}^\infty\exp\{-c\Lambda_t\}<\infty$), Lemma C.10 immediately implies that $\left|\mathsf{R}_{t_1,t_2}-\tau^{\star 2}\right|\to 0$ as $t_1,t_2\ge t_0$ and $t_0\to\infty$. With the assistance of the aforementioned two lemmas, we are ready to proceed to the proofs of Lemma C.5 and Lemma C.7.

C.4.2 Proof of Lemma C.5

Now we are ready to prove Lemma C.5. It suffices to show that $\lim_{p\to\infty}|S_{t_2}(\gamma)\setminus S_{t_1}(\gamma)|/p\le\xi$ almost surely. As argued in the proof of BM-Lemma 3.5, it is guaranteed that

$$\lim_{p\to\infty}\frac{1}{p}|S_{t_2}(\gamma)\setminus S_{t_1}(\gamma)| = \lim_{p\to\infty}\mathbb{P}\left\{|\Theta+Z_{t_2-1}|\ge(1-\gamma)\zeta_{t_2-1}, |\Theta+Z_{t_1-1}|<(1-\gamma)\zeta_{t_1-1}\right\} =: P_{t_1,t_2},$$

where (Z_{t_1},Z_{t_2}) are jointly Gaussian with $\mathbb{E}[Z_{t_1}^2]=\mathsf{R}_{t_1,t_1}$, $\mathbb{E}[Z_{t_2}^2]=\mathsf{R}_{t_2,t_2}$ and $\mathbb{E}[Z_{t_1}Z_{t_2}]=\mathsf{R}_{t_1,t_2}$. If we denote $a\coloneqq(1-\gamma)\alpha^\star\tau^\star$, in view of Lemma B.3, we know that $\forall\varepsilon>0$ and large enough $t_\star$, it holds that $|(1-\gamma)\zeta_{t_i-1}-a|\le\varepsilon$ for $i\in\{1,2\}$. Then with the same argument as BM-Lemma 3.5, we reach

$$\begin{aligned}
P_{t_1,t_2} &\le \frac{1}{4\varepsilon^2}\left[\mathsf{R}_{t_1-1,t_1-1}-2\mathsf{R}_{t_1-1,t_2-1}+\mathsf{R}_{t_2-1,t_2-1}\right]+\frac{4\varepsilon}{\sqrt{2\pi\mathsf{R}_{t_1-1,t_1-1}}}\\
&\le \frac{\left[\sum_{i=t_\star}^\infty\sqrt{c_0\exp\{-c\Lambda_i\}}+\sum_{i=t_\star}^\infty\sqrt{4L_\tau\tau^\star_{\max}l_i}\right]^2}{\varepsilon^2}+\frac{2\varepsilon}{\sigma},
\end{aligned}$$

as a consequence of Lemma C.10 and $\tau_t\ge\sigma$, $\forall t\ge 1$.

Under Assumption 1, $\sum_{i=t_\star}^\infty\sqrt{l_i}\to 0$ and $\sum_{i=t_\star}^\infty\exp\{-c/2\Lambda_i\}\to 0$ as $t_\star$ increases. Taking $\varepsilon=\left[\sum_{i=t_\star}^\infty\sqrt{c_0\exp\{-c\Lambda_i\}}+\sum_{i=t_\star}^\infty\sqrt{4L_\tau\tau^\star_{\max}l_i}\right]^{2/3}$ gives

$$P_{t_1,t_2}\le C'\left[\sum_{i=t_\star}^\infty\sqrt{c_0\exp\{-c\Lambda_i\}}+\sum_{i=t_\star}^\infty\sqrt{4L_\tau\tau^\star_{\max}l_i}\right]^{2/3}.$$

The conclusion of Lemma C.5 thus follows immediately.

C.4.3 Proof of Lemma C.7

In view of the proof for BM-Lemma 4.3 — a general result proved for any positive thresholding sequence $\{\zeta_t\}$), we obtain

$$\lim_{p\to\infty}\frac{1}{p\lambda_t^2}\|\boldsymbol{z}^t-\boldsymbol{z}^{t-1}\|_2^2 \overset{\text{a.s.}}{=} \lim_{p\to\infty}\frac{1}{p\lambda_t^2}\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\|_2^2.$$

It is thus sufficient to prove Lemma C.7 for the $\boldsymbol{\theta}^t$ sequence only. As a consequence of the generalized state evolution formula (cf. Lemma C.8), we can guarantee that

$$\lim_{p\to\infty}\frac{1}{p\lambda_t^2}\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\|_2^2=\frac{1}{\lambda_t^2}\mathbb{E}\left\{[\eta(\Theta+Z_t;\zeta_t)-\eta(\Theta+Z_{t-1};\zeta_{t-1})]^2\right\}$$
$$\leq\frac{2(\zeta_t-\zeta_{t-1})^2}{\lambda_t^2}+\frac{2\mathbb{E}[(Z_t-Z_{t-1})^2]}{\lambda_t^2}.$$

- The term $\mathbb{E}[(Z_t-Z_{t-1})^2]$ can be controlled via Lemma C.9, where

$$\mathbb{E}[(Z_t-Z_{t-1})^2]=\mathsf{R}_{t,t}-2\mathsf{R}_{t-1,t}+\mathsf{R}_{t-1,t-1}\leq c_0\exp\{-c\Lambda_t\}+4L\tau_{\max}^\star l_t. \tag{77}$$

  To control the right-hand side, we first obtain from Assumption 1 that $\sum_{s=t}^\infty\sqrt{l_s}$ is a converging sequence. In addition, Assumption 1 ensures that $\sum_{s=t}^\infty\lambda_s$ is a diverging sequence. Then we conclude that $l_{t-1}/\lambda_t^2\to 0$ as $t\to\infty$. Additionally, by Assumption 1, we know that $\sum_{s=1}^\infty\exp\{-c/2\Lambda_t\}$ diverges, thus $\exp\{-c\Lambda_t\}/\lambda_t^2\to 0$. Taking these collectively ensures that

$$\frac{2\mathbb{E}[(Z_t-Z_{t-1})^2]}{\lambda_t^2}\to 0.$$

- It remains to show that the term $2(\zeta_t-\zeta_{t-1})^2/\lambda_t^2$ vanishes as $t\to\infty$. We obtain from Lemma B.2 that $0\leq\alpha^{\star\prime}(0)\leq C$. It then immediately follows that

$$\left|\frac{\zeta_t-\zeta_{t-1}}{\lambda_t}\right|\leq\alpha_t^\star\left|\frac{\tau_t-\tau_{t-1}}{\lambda_t}\right|+\tau_{t-1}\left|\frac{\alpha_t^\star-\alpha_{t-1}^\star}{\lambda_t-\lambda_{t-1}}\right|\frac{\lambda_t-\lambda_{t-1}}{\lambda_t}\to\alpha^\star\left|\frac{\tau_t-\tau_{t-1}}{\lambda_t}\right|+\tau^\star\alpha^{\star\prime}(0)\frac{\lambda_t-\lambda_{t-1}}{\lambda_t}\to 0,$$

  where the limit value $|\tau_t-\tau_{t-1}|/\lambda_t\overset{\text{a.s.}}{\to}0$ follows from Lemma C.9, and $(\lambda_t-\lambda_{t-1})/\lambda_t\to 0$ holds by virtue of Assumption 1.

Putting all this together completes the proof.

C.4.4 Proof of Lemma C.10

The main idea of this proof is to decompose $\mathsf{R}_{t_1,t_2}-\tau^{\star 2}$ into terms of the form $\mathsf{R}_{t,t+1}-\tau^{\star 2}$ or $\mathsf{R}_{t,t}-\tau^{\star 2}$. Without loss of generality, we assume $t_2>t_1$. The proof of BM-Theorem 4.2 (which is a general result for any positive $\{\zeta_t\}$ sequence, so we can safely use the arguments there) ensures that: if we define

$$\boldsymbol{h}^{t+1}=\boldsymbol{\theta}^\star-(\boldsymbol{X}^\top\boldsymbol{z}^t+\boldsymbol{\theta}^t),$$

then the empirical distribution of $\{\boldsymbol{h}^{i+1}\}_{T\geq i\geq 0}$ converges weakly to a sequence of Gaussian random variables $\{Z_i\}_{T\geq i\geq 0}$ as $p\to\infty$. Here, $Z_0,Z_1,Z_2,\dots$ are Gaussian random variables, defined on the same probability space with $\mathbb{E}[Z_t]=0$ and $\mathbb{E}[Z_tZ_s]=\mathsf{R}_{t,s}$. In addition, one has

$$\mathsf{R}_{t_1,t_2}=\lim_{p\to\infty}\frac{1}{p}\langle\boldsymbol{h}^{t_1+1},\boldsymbol{h}^{t_2+1}\rangle.$$

With these results in place, direct calculations yield

$$|\mathsf{R}_{t_1,t_1}-2\mathsf{R}_{t_1,t_2}+\mathsf{R}_{t_2,t_2}|=\mathbb{E}[(Z_{t_1}-Z_{t_2})^2]=\sum_{i,j=t_1}^{t_2-1}\mathbb{E}[(Z_{i+1}-Z_i)(Z_{j+1}-Z_j)]$$
$$\leq\left[\sum_{i=t_1}^{t_2-1}\left\{\mathbb{E}(Z_{i+1}-Z_i)^2\right\}^{1/2}\right]^2$$
$$\leq 4\left[\sum_{i=t_1}^{\infty}\sqrt{c_0\exp\{-c\Lambda_i\}+4L_\tau\tau_{\max}^\star l_i}\right]^2,$$

where the last inequality follows from inequality (77). The proof of Lemma C.10 is thus completed.

C.4.5 Proof of Lemma C.9

Similar to the proof of BM-Lemma 5.7, we find it convenient to change coordinates and define

$$y_{t,1} \coloneqq \mathsf{R}_{t-1,t-1} = \tau_{t-1}^2, \quad y_{t,2} \coloneqq \mathsf{R}_{t,t} = \tau_t^2, \quad y_{t,3} \coloneqq \mathsf{R}_{t-1,t-1} - 2\mathsf{R}_{t-1,t} + \mathsf{R}_{t,t}.$$

To capture the updating rule (73), we introduce the following recursive formula via the mapping $\boldsymbol{y}_{t+1} = \mathsf{G}^t(\boldsymbol{y}_t)$:

$$y_{t+1,1} = \mathsf{G}_1^t(\boldsymbol{y}_t) \coloneqq y_{t,2}, \tag{78a}$$

$$y_{t+1,2} = \mathsf{G}_2^t(\boldsymbol{y}_t) \coloneqq \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{\left[\eta(\Theta + Z_t; \alpha_t^\star\sqrt{y_{t,2}}) - \Theta\right]^2\right\} = \mathsf{F}(y_{t,2}, \alpha_t^\star\sqrt{y_{t,2}}), \tag{78b}$$

$$y_{t+1,3} = \mathsf{G}_3^t(\boldsymbol{y}_t) \coloneqq \frac{1}{\delta}\mathbb{E}\left\{\left[\eta(\Theta + Z_t; \alpha_t^\star\sqrt{y_{t,2}}) - \eta(\Theta + Z_{t-1}; \alpha_{t-1}^\star\sqrt{y_{t,1}})\right]^2\right\}, \tag{78c}$$

where (Z_t, Z_{t-1}) are jointly mean-zero Gaussian random variables with $\mathbb{E}[Z_t^2] = y_{t,2}$, $\mathbb{E}[Z_{t-1}^2] = y_{t,1}$ and $\mathbb{E}[(Z_t - Z_{t-1})^2] = y_{t,3}$. In contrast to BM-Lemma 5.7, the mapping G^t is different for each step t since at different t, $\alpha_t^\star$ is chosen as the unique solution to the equation set (26), which varies across iterations. (As such, the mapping G^t is well-defined for $y_{t,3} \le 2(y_{t,1} + y_{t,2})$.) In addition, the recursive updates are initialized to $Z_0 \sim \mathcal{N}(0,1)$ and

$$\begin{aligned}
y_{1,1} &\coloneqq \sigma^2 + \frac{1}{\delta}\mathbb{E}[\Theta^2];\\
y_{1,2} &\coloneqq \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{\left[\eta(\Theta + Z_0; \zeta_0) - \Theta\right]^2\right\};\\
y_{1,3} &\coloneqq \frac{1}{\delta}\mathbb{E}\left\{\eta(\Theta + Z_0; \zeta_0)^2\right\}.
\end{aligned}$$

We first make note of the following fact: if $y_{t,1} = y_{t,2} = \tau_t^{\star 2}$, then one has $y_{t+1,1} = y_{t+1,2} = \tau_t^{\star 2}$ since $\alpha_t^\star$ satisfies the equation set (26) where

$$\tau_t^{\star 2} = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{\left[\eta(\Theta + \tau_t^\star Z_t; \alpha_t^\star\tau_t^\star) - \Theta\right]^2\right\}.$$

In other words, $\boldsymbol{y}_t^\star$ is a fixed point of the mapping G^t, namely, $\boldsymbol{y}_t^\star = \mathsf{G}^t(\boldsymbol{y}_t^\star)$. In addition, one has $\lim_{t\to\infty}\tau_t^\star = \tau^\star$ by Proposition B.3.

We claim that the following three properties hold true.

1. As $t \to \infty$, the update sequence satisfies $y_{t,1} \to \tau_t^{\star 2}$ and $y_{t,2} \to \tau_t^{\star 2}$, with $y_{t,3} < y_{t,1} + y_{t,2} - \sigma^2$.
2. Define another recursive updating rule $\widetilde{y}_{t+1,3} \coloneqq \mathsf{G}_3^t(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3})$. If we can find some $t_0 \ge T_{\min}$ such that $\widetilde{y}_{t_0,3} < 2\tau_{t_0}^{\star\,2}$, then the following two properties are satisfied
   (a) for all $t \ge t_0$, it holds that $\widetilde{y}_{t,3} < 2\tau_t^{\star 2}$;
   (b) it follows that $\widetilde{y}_{t,3} \to 0$ as $t \to \infty$.

   Here $T_{\min}$ is some constant that is pre-determined by Θ and σ.
3. When $t \ge t_0$, the Jacobian $J_t \coloneqq J_{G^t}(\boldsymbol{y}_t^\star)$ of G^t at $\boldsymbol{y}_t^\star \coloneqq (\tau_t^{\star 2}, \tau_t^{\star 2}, 0)$ has spectral radius

$$\sigma(J_t) \le 1 - \frac{\lambda_t}{2\alpha_{\max}^\star\tau_{\max}^\star}. \tag{79}$$

Let us take these claims as given for the moment and proceed to the proof of Lemma C.9. In light of the first claim, we obtain that for all large enough t, one has $y_{t,3} \le 2\tau_t^{\star 2} - \sigma^2$. Then the second claim further guarantees that $y_{t,3} \to 0$. Taking the first two claims collectively implies that $\boldsymbol{y}_t \to \boldsymbol{y}_t^\star = (\tau_t^{\star 2}, \tau_t^{\star 2}, 0)$ as $t \to \infty$. In addition, by virtue of the third property, for appropriately large t, we obtain

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^\star\|_2 = \|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t^\star\|_2 + \|\boldsymbol{y}_{t+1}^\star - \boldsymbol{y}_t^\star\|_2 \overset{\text{(i)}}{\le} \sigma(J_t)\|\boldsymbol{y}_t - \boldsymbol{y}_t^\star\|_2 + \|\boldsymbol{y}_{t+1}^\star - \boldsymbol{y}_t^\star\|_2$$

$$\overset{\text{(ii)}}{\leq} \exp\left\{-c\lambda_t\right\} \|\boldsymbol{y}_t - \boldsymbol{y}_t^\star\|_2 + 4\tau_{\max}^\star |\tau_t^\star - \tau_{t+1}^\star|, \tag{80}$$

where we define $c = (2\alpha_{\max}^\star \tau_{\max}^\star)^{-1}$. Here, the inequality (i) uses the relations

$$\boldsymbol{y}_{t+1} = \mathsf{G}^t(\boldsymbol{y}_t) \quad \text{and} \quad \boldsymbol{y}_t^\star = \mathsf{G}^t(\boldsymbol{y}_t^\star),$$

and (ii) uses the property (79). Recalling that we define $\Lambda_t \coloneqq \sum_{i \leq t} \lambda_i$, we can apply the inequality (80) recursively to yield

$$\begin{aligned}
\frac{\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^\star\|_2}{\exp\left\{-c\Lambda_t\right\}} &\leq \frac{\|\boldsymbol{y}_t - \boldsymbol{y}_t^\star\|_2}{\exp\left\{-c\Lambda_{t-1}\right\}} + 4\tau_{\max}^\star \frac{|\tau_t^\star - \tau_{t+1}^\star|}{\exp\left\{-c\Lambda_t\right\}} \\
&\leq \frac{\|\boldsymbol{y}_{t_0} - \boldsymbol{y}_{t_0}^\star\|_2}{\exp\left\{-c\Lambda_{t_0-1}\right\}} + 4\tau_{\max}^\star \sum_{s=t_0}^{t} \frac{|\tau_t^\star - \tau_{s+1}^\star|}{\exp\left\{-c\Lambda_s\right\}} \\
&\leq \frac{\|\boldsymbol{y}_{t_0} - \boldsymbol{y}_{t_0}^\star\|_2}{\exp\left\{-c\Lambda_{t_0-1}\right\}} + 4L_\tau \tau_{\max}^\star \sum_{s=1}^{t} \frac{|\lambda_s - \lambda_{s+1}|}{\exp\left\{-c\Lambda_s\right\}}.
\end{aligned}$$

Then we can conclude that, with $c_0 \coloneqq \|\boldsymbol{y}_{t_0} - \boldsymbol{y}_{t_0}^\star\|_2 \exp\left\{c\Lambda_{t_0-1}\right\}$,

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^\star\|_2 \leq c_0 \exp\left\{-c\Lambda_t\right\} + 4L\tau_{\max}^\star \sum_{s=1}^{t} |\lambda_s - \lambda_{s+1}| \exp\left\{-c[\Lambda_t - \Lambda_s]\right\}.$$

Finally, recalling the definition

$$y_{t,1} = \mathsf{R}_{t-1,t-1} = \tau_{t-1}^2, \quad y_{t,2} = \mathsf{R}_{t,t} = \tau_t^2, \quad y_{t,3} = \mathsf{R}_{t-1,t-1} - 2\mathsf{R}_{t-1,t} + \mathsf{R}_{t,t},$$

we can show that

$$\max\left\{|\mathsf{R}_{t,t} - \tau_t^{\star 2}|, |\mathsf{R}_{t,t} - 2\mathsf{R}_{t,t+1} + \mathsf{R}_{t+1,t+1}|\right\} \leq c_0 \exp\left\{-c\Lambda_t\right\} + 4L\tau_{\max}^\star \sum_{s=1}^{t} |\lambda_s - \lambda_{s+1}| \exp\left\{-c[\Lambda_t - \Lambda_s]\right\},$$

which establishes the advertised result in Lemma C.9. It remains to prove the three claims above.

C.4.6 Proof of the three claims

In the sequel, we look at these three claims separately.

Proof of Claim 1

The iteration updates (78a) and (78b), along with the initial condition, yield that $y_{t+1,1} = y_{t,2} = \tau_t$, $\forall t \geq 1$. Taking the results in Proposition B.3 collectively with $\lim_{t\to\infty} \tau_t = \tau^\star$ and $\lim_{t\to\infty} \tau_t^\star = \tau^\star$, we obtain

$$y_{t,1} - \tau_t^{\star 2} \to 0 \qquad \text{and} \qquad y_{t,2} - \tau_t^{\star 2} \to 0.$$

The second part of the claim is proved by induction; for the proof to go through, let us verify the initial condition and the induction step respectively.

- For the initial condition, we can write

$$y_{1,1} + y_{1,2} - y_{1,3} = 2\sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{\Theta\left(\Theta - \eta(\Theta + Z_0; \zeta_0)\right)\right\}.$$

  It is easy to check that the function $\theta \mapsto \theta - \eta(\theta + Z_0; \zeta_0)$ is monotonically increasing for any Z_0 and ζ_0; therefore, the random variables Θ and $\Theta - \eta(\Theta + Z_0; \zeta_0)$ are positively correlated. In other words,

$$\frac{1}{\delta}\mathbb{E}\left\{\Theta\left(\Theta - \eta(\Theta + Z_0; \zeta_0)\right)\right\} \geq 0 \quad \Longrightarrow \quad y_{1,1} + y_{1,2} - y_{1,3} \geq 2\sigma^2.$$

- Suppose $y_{t,3} < y_{t,1} + y_{t,2} - \sigma^2$ holds for all steps up to t. At step $t+1$, the iteration formula (78) directly leads to

$$y_{t+1,1} + y_{t+1,2} - y_{t+1,3} = 2\sigma^2 + \frac{2}{\delta}\mathbb{E}\left[\left(\eta(\Theta + Z_t; \alpha_t^\star\sqrt{y_{t,2}}) - \Theta\right)\left(\eta(\Theta + Z_{t-1}; \alpha_{t-1}^\star\sqrt{y_{t,1}}) - \Theta\right)\right],$$

where (Z_t, Z_{t-1}) are jointly Gaussian with $\mathbb{E}[Z_t^2] = y_{t,2}$, $\mathbb{E}[Z_{t-1}^2] = y_{t,1}$ and $\mathbb{E}[(Z_t - Z_{t-1})^2] = y_{t,3}$. By definition of Z_t and Z_{t-1}, one has $\mathbb{E}[Z_t Z_{t-1}] = (y_{t,1} + y_{t,2} - y_{t,3})/2$ which stays positive given the induction assumption.

Now since the mapping $x \mapsto \eta(x + \theta; \zeta) - \theta$ is monotone in x, we can readily conclude that

$$\mathbb{E}\left[\left(\eta(\Theta + Z_t; \alpha_t^\star\sqrt{y_{t,2}}) - \Theta\right)\left(\eta(\Theta + Z_{t-1}; \alpha_{t-1}^\star\sqrt{y_{t,1}}) - \Theta\right)\right] \geq 0,$$

which further leads to $y_{t+1,1}+y_{t+1,2}-y_{t+1,3} > 2\sigma^2$. We have thus proved that $y_{t+1,3} < y_{t+1,1}+y_{t+1,2}-\sigma^2$ at $t+1$.

Combining the initial condition with the induction argument, we conclude that $y_{t,3} < y_{t,1} + y_{t,2} - \sigma^2$ for all t.

Proof of Claim 2(a)

To establish the second claim, with a slight abuse of notation, we define

$$G_3(y_1, y_2, y_3; \alpha_1, \alpha_2) \coloneqq \frac{1}{\delta}\mathbb{E}\left\{[\eta(\Theta + Z_1; \alpha_1\sqrt{y_1}) - \eta(\Theta + Z_2; \alpha_2\sqrt{y_2})]^2\right\}, \tag{81}$$

where (Z_1, Z_2) are jointly Gaussian with zero mean and $\mathbb{E}[Z_1^2] = y_1$, $\mathbb{E}[Z_2^2] = y_2$ and $\mathbb{E}[(Z_1 - Z_2)^2] = y_3$. We remark that the dependence on y_3 is only through the covariance between Z_1 and Z_2. From the definition of $\widetilde{y}_{t+1,3}$, we immediately have

$$\widetilde{y}_{t+1,3} = G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}, \alpha_t^\star, \alpha_{t-1}^\star).$$

It is straightforward to verify that G_3 is continuous in all the parameters under our assumption that $\mathbb{E}[\Theta^2] < \infty$. The continuity of G_3, together with the fact that $\alpha_t^{\star 2}$ converges, implies that $\widetilde{y}_{t+1,3} - G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) \to 0$, for any fixed $t \geq 0$.

With this notation in mind, we are ready to prove that $\widetilde{y}_{t+1,3} < 2(\tau_{t+1}^\star)^2$ for all sufficiently large t via an inductive argument. First, suppose $\widetilde{y}_{i,3} < 2(\tau_i^\star)^2$ holds for all steps between t_0 up to t. In view of relation $\tau_t^{\star 2} = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta(\Theta + Z_t; \alpha_t^\star\tau_t^\star) - \Theta]^2\right\}$, we arrive at

$$2\tau_t^{\star 2} - G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) = 2\sigma^2 + \frac{2}{\delta}\mathbb{E}\left\{\left(\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) - \Theta\right)\left(\eta(\Theta + Z_2; \alpha_t^\star\tau_t^\star) - \Theta\right)\right\},$$

where again (Z_1, Z_2) are jointly Gaussian with zero mean and $\mathbb{E}[Z_1^2] = \tau_t^{\star 2}$, $\mathbb{E}[Z_2^2] = \tau_t^{\star 2}$ and $\mathbb{E}[Z_1 Z_2] = 2\tau_t^{\star 2} - \widetilde{y}_{t,3}$. Under the induction assumption $\widetilde{y}_{t,3} \leq 2\tau_t^{\star 2}$, the two random variables $\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) - \Theta$ and $\eta(\Theta + Z_2; \alpha_t^\star\tau_t^\star) - \Theta$ are positively or zero correlated. Then it is guaranteed that

$$\frac{2}{\delta}\mathbb{E}\left\{\left(\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) - \Theta\right)\left(\eta(\Theta + Z_2; \alpha_t^\star\tau_t^\star) - \Theta\right)\right\} \geq 0 \quad \Longrightarrow \quad 2\tau_t^{\star 2} - G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) \geq 2\sigma^2. \tag{82}$$

Now to prove the upper bound for $\widetilde{y}_{t+1,3}$, it is sufficient for us to control the difference between $G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star)$ and $\widetilde{y}_{t+1,3}$. Towards this end, let us invoke the Lipschitz property for soft-thresholding function

$$\left|\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) - \eta(\Theta + Z_1; \alpha_{t-1}^\star\tau_t^\star)\right| \leq \tau_t^\star|\alpha_t^\star - \alpha_{t-1}^\star|,$$

which leads to

$$
\begin{aligned}
&|G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) - \widetilde{y}_{t+1,3}| \\
=&|G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) - G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_{t-1}^\star, \alpha_t^\star)| \\
=& \frac{1}{\delta}\mathbb{E}\left|\left[\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) - \eta(\Theta + Z_1; \alpha_{t-1}^\star\tau_t^\star)\right]\left[\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) + \eta(\Theta + Z_1; \alpha_{t-1}^\star\tau_t^\star) - 2\eta(\Theta + Z_2; \alpha_t^\star\tau_t^\star)\right]\right| \\
\le& \frac{\tau_t^\star|\alpha_t^\star - \alpha_{t-1}^\star|}{\delta}\mathbb{E}\left|\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) + \eta(\Theta + Z_1; \alpha_{t-1}^\star\tau_t^\star) - 2\eta(\Theta + Z_2; \alpha_t^\star\tau_t^\star)\right| \\
\overset{\text{(i)}}{\le}& \frac{\tau_t^\star|\alpha_t^\star - \alpha_{t-1}^\star|}{\delta} \cdot 4(\mathbb{E}[|\Theta|] + \tau_{\max}^\star) \le \frac{4(\mathbb{E}[|\Theta|] + \tau_{\max}^\star)\tau_{\max}^\star}{\delta}|\alpha_t^\star - \alpha_{t-1}^\star|.
\end{aligned} \tag{83}
$$

Here, the inequality (i) follows since $\mathbb{E}\left[|\eta(\Theta + Z_i; \alpha_t^\star\tau_s^\star)|\right] \le \mathbb{E}\left[|\Theta + Z_1|\right] \le \mathbb{E}[|\Theta|] + \tau_{\max}^\star$ with $i \in \{1, 2\}$ and $s \in \{t, t-1\}$.

Combining the inequalities (82) and (83) yields

$$
\widetilde{y}_{t+1,3} \le 2{\tau_{t+1}^\star}^2 + \frac{4(\mathbb{E}[|\Theta|] + \tau_{\max}^\star)\tau_{\max}^\star}{\delta}|\alpha_t^\star - \alpha_{t-1}^\star| + 2(\tau_t^{\star 2} - {\tau_{t+1}^\star}^2) - 2\sigma^2.
$$

As $t \to \infty$, we know from Proposition B.3 that $|\alpha_t^\star - \alpha_{t-1}^\star| \to 0$, as well as $|\tau_t^{\star 2} - {\tau_{t+1}^\star}^2| \to 0$; as such, we can always find some $T_{\min}$ determined by Θ and σ^2, such that the $\frac{4(\mathbb{E}[|\Theta|] + \tau_{\max}^\star)\tau_{\max}^\star}{\delta}|\alpha_t^\star - \alpha_{t-1}^\star| + 2|\tau_t^{\star 2} - {\tau_{t+1}^\star}^2| - 2\sigma^2 < 0$. In this case, we conclude that $\widetilde{y}_{t+1,3} < 2{\tau_{t+1}^\star}^2$, and then by induction, $\widetilde{y}_{t,3} < 2{\tau_{t+1}^\star}^2$, $\forall t \ge t_0$. This completes the proof of Claim 2(a).

Proof of Claim 2(b)

It turns out that the function $G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star)$ is the same as the G_* function defined in BM-Pg 36. We record here two key observations from BM-Pg 36 regarding this function.

- The derivative of function G_3 with respect to its third argument satisfies

$$
\frac{\partial}{\partial x}G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, x; \alpha_t^\star, \alpha_t^\star)\Big|_{x=0} = \frac{1}{\delta}\mathbb{E}\left\{\eta'(\Theta + \tau_t^\star Z; \alpha_t^\star\tau_t^\star)\right\} = 1 - \frac{\lambda_t}{\alpha_t^\star\tau_t^\star}; \tag{84}
$$

- The mapping $x \mapsto \frac{\partial}{\partial x}G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, x; \alpha_t^\star, \alpha_t^\star)$ is decreasing in $[0, 2\tau_t^{\star 2})$.

In addition, we claim that $G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, 0; \alpha_t^\star, \alpha_t^\star) = 0$. In order to see this, by construction in (81), we have

$$
G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, 0; \alpha_t^\star, \alpha_t^\star) = \frac{1}{\delta}\mathbb{E}\left\{\left[\eta(\Theta + Z_1; \alpha_t^\star\tau_t^\star) - \eta(\Theta + Z_2; \alpha_t^\star\tau_t^\star)\right]^2\right\},
$$

where $\mathbb{E}[Z_1^2] = \mathbb{E}[Z_2^2] = \tau_t^{\star 2}$ and $\mathbb{E}[(Z_1 - Z_2)^2] = 0$. From this construction, we know $Z_1 = Z_2$, almost surely; it follows that $\eta(\Theta + Z_1; \alpha_1\tau_t^\star) = \eta(\Theta + Z_1; \alpha_1\tau_t^\star)$, almost surely. We therefore obtain $G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, 0; \alpha_t^\star, \alpha_t^\star) = 0$.

Taking these observations collectively guarantees that: under the condition $\widetilde{y}_{t,3} < 2\tau_t^{\star 2}$, one has

$$
G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) \le (1 - \lambda_t/\alpha_t^\star\tau_t^\star)\widetilde{y}_{t,3} \le (1 - (\alpha_{\max}^\star\tau_{\max}^\star)^{-1}\lambda_t)\widetilde{y}_{t,3}.
$$

In view of the inequality (83), we have

$$
|G_3(\tau_t^{\star 2}, \tau_t^{\star 2}, \widetilde{y}_{t,3}; \alpha_t^\star, \alpha_t^\star) - \widetilde{y}_{t+1,3}| \le C|\alpha_t^\star - \alpha_{t-1}^\star|,
$$

where C is determined by Θ and σ^2. Putting these pieces together leads to

$$
\widetilde{y}_{t+1,3} \le (1 - (\alpha_{\max}^\star\tau_{\max}^\star)^{-1}\lambda_t)\widetilde{y}_{t,3} + C|\alpha_{t+1}^\star - \alpha_t^\star|. \tag{85}
$$

Invoking the above relation recursively, we obtain, for $c_1 \coloneqq (\alpha_{\max}^\star\tau_{\max}^\star)^{-1}$, that

$$
\widetilde{y}_{t+1,3} \le \prod_{i=0}^{t-t_0}(1 - c_1\lambda_{t-i})\widetilde{y}_{t_0} + C\sum_{i=0}^{t-t_0}\prod_{j=1}^{i}(1 - c_1\lambda_{t+1-j})|\alpha_{t+1-i}^\star - \alpha_{t-i}^\star|
$$

$$\leq e^{-c_1\sum_{i=t_0}^{t}\lambda_i}\widetilde{y}_{t_0} + C\sum_{i=t_0}^{t} e^{-c_1\sum_{j=i+1}^{t}\lambda_j}|\alpha_{i+1}^\star - \alpha_i^\star|$$

$$= e^{-c_1\sum_{i=t_0}^{t}\lambda_i}\widetilde{y}_{t_0} + C\sum_{i=t_0}^{t/2} e^{-c_1\sum_{j=i+1}^{t}\lambda_j}|\alpha_{i+1}^\star - \alpha_i^\star| + C\sum_{i=t/2}^{t} e^{-c_1\sum_{j=i+1}^{t}\lambda_j}|\alpha_{i+1}^\star - \alpha_i^\star|. \tag{86}$$

Observing that $\widetilde{y}_{t_0,3} < 2{\tau_{t_0}^\star}^2$ and the sequence $\sum_{i=t_0}^{t}\lambda_i$ diverges to infinity, we know that the first term on the right-hand side of (86) converges to zero as t increases. Moreover, as proved in Proposition B.2, the sequence $\{\alpha_t^\star\}$ is monotone when t is sufficiently large, and we then see that $\sum_{t=t/2}^{+\infty}|\alpha_{t+1}^\star - \alpha_t^\star|$ converges to zero. In the meantime, $e^{-c_1\sum_{j=i+1}^{t}\lambda_j} \leq e^{-c_1\lambda_t} \leq e^{-c_1\max_t\lambda_t}$ is always controlled by some constant. As a result, the third term on the right-hand side of (86) also vanishes. It remains to control the second term on the right-hand side of (86). To this end, we make the observation that $|\alpha_{i+1}^\star - \alpha_i^\star| \leq 2\alpha_{\max}^\star$ and

$$\sum_{i=t_0}^{t/2} e^{-c_1\sum_{j=i+1}^{t}\lambda_j} \leq te^{-c_1\sum_{j=t/2}^{t}\lambda_j} = e^{(\log t - c_1\sum_{j=t/2}^{t}\lambda_j)} \to 0,$$

where the last inequality follows from Assumption 1. Thus, the second term on the right-hand side of (86) also has a zero limit. Taking these collectively, we establish the advertised result in Claim 2(b).

Proof of Claim 3.

To begin with, by some direct algebra, we can express the Jacobian matrix of mapping G^t at $\boldsymbol{y}_\star = ({\tau_t^\star}^2, {\tau_t^\star}^2, 0)$ by

$$J_{\mathsf{G}^t}(\boldsymbol{y}_\star) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{F}(\tau^2, \alpha_t^\star\tau)\big|_{{\tau_t^\star}^2} & 0 \\ \cdots & \cdots & \frac{\mathrm{d}}{\mathrm{d}y_3}\mathsf{G}_3^t({\tau_t^\star}^2, {\tau_t^\star}^2, y_3)\Big|_{y_3=0} \end{pmatrix}.$$

It is easily seen that the maximal eigenvalues of the above matrix satisfies the following relation

$$\sigma(J_{\mathsf{G}^t}(\boldsymbol{y}_\star)) = \max\left\{ \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{F}(\tau^2, \alpha_t^\star\tau)\Big|_{{\tau_t^\star}^2}, \frac{\mathrm{d}}{\mathrm{d}y_3}\mathsf{G}_3^t({\tau_t^\star}^2, {\tau_t^\star}^2, y_3)\Big|_{y_3=0} \right\}. \tag{87}$$

To control $\sigma(J_{\mathsf{G}^t}(\boldsymbol{y}_\star))$, it is sufficient for us to bound each term respectively. For the first term, from the proof of BM-Proposition 1.3, we know that the function $\tau^2 \mapsto \mathsf{F}(\tau^2, \alpha\tau)$ is concave, for any $\alpha > 0$ and Θ that is not identically 0; therefore, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{F}(\tau^2, \alpha_t^\star\tau)\Big|_{{\tau_t^\star}^2} \leq \frac{\mathsf{F}({\tau_t^\star}^2, \alpha_t^\star\tau_t^\star) - \mathsf{F}(0,0)}{{\tau_t^\star}^2 - 0} = \frac{{\tau_t^\star}^2 - \sigma^2}{{\tau_t^\star}^2} \leq 1 - \frac{\sigma^2}{{\tau_{\max}^\star}^2} \coloneqq \eta.$$

For the second term, with the function G_3 defined in expression (81), we write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y_3}\mathsf{G}_3^t({\tau_t^\star}^2, {\tau_t^\star}^2, y_3)\Big|_{y_3=0} &= \frac{\mathrm{d}}{\mathrm{d}x}G_3({\tau_t^\star}^2, {\tau_t^\star}^2, x; \alpha_{t-1}^\star, \alpha_t^\star)\Big|_{x=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left[\eta(\Theta + a(x)Z + b(x)W; \alpha_t^\star\tau_t^\star)\eta(\Theta + \tau_t^\star Z; \alpha_{t-1}^\star\tau_t^\star)\right]\Big|_{x=0}.\end{aligned}$$

Here Z and W are independent standard Gaussian random variables and we denote

$$a(x) \coloneqq \tau_t^\star - x/(2\tau_t^\star) \qquad \text{and} \quad b(x) \coloneqq \sqrt{x - x^2/(4{\tau_t^\star}^2)}. \tag{88}$$

Further define

$$g(\Theta, Z, W; x) \coloneqq \eta(\Theta + a(x)Z + b(x)W; \alpha_t^\star\tau_t^\star)\Big[\eta(\Theta + \tau_t^\star Z; \alpha_{t-1}^\star\tau^\star) - \eta(\Theta + \tau^\star Z; \alpha_t^\star\tau_t^\star)\Big]. \tag{89}$$

With this notation in place, we can write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y_3}\mathsf{G}_3^t(\tau_t^{\star 2},\tau_t^{\star 2},y_3)\Big|_{y_3=0} &= \frac{\mathrm{d}}{\mathrm{d}y_3}G_3(\tau_t^{\star 2},\tau_t^{\star 2},x;\alpha_t^\star,\alpha_t^\star)\Big|_{x=0} + \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left\{g(\Theta,Z,W;x)\right\}\Big|_{x=0} \\ &= 1-\frac{\lambda_t}{\tau_t^\star\alpha_t^\star} + \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left\{g(\Theta,Z,W;x)\right\}\Big|_{x=0},\end{aligned} \tag{90}$$

where the last equality follows from the expression (84). To control the spectral radius of the Jacobian, it suffices to control the last term on the right-hand side above. We claim that it satisfies the following property:

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left\{g(\Theta,Z,W;x)\right\}\Big|_{x=0} \le 2|\alpha_t^\star-\alpha_{t-1}^\star|. \tag{91}$$

Taking this claim as given for the moment, we can translate the expression (90) into

$$\frac{\mathrm{d}}{\mathrm{d}y_3}\mathsf{G}_3^t(\tau_t^{\star 2},\tau_t^{\star 2},y_3)\Big|_{y_3=0} \le 1-\frac{\lambda_t}{\alpha_{\max}^\star\tau_{\max}^\star}+2|\alpha_t^\star-\alpha_{t-1}^\star|.$$

Combining the results of Proposition B.2 and Assumption 1 gives

$$|\alpha_t^\star-\alpha_{t-1}^\star| \le L_\alpha|\lambda_t-\lambda_{t-1}| \le \frac{\lambda_t}{4\alpha_{\max}^\star\tau_{\max}^\star}$$

for sufficiently large t. Thus we complete the proof of expression (79). The only thing that is left is to establish the inequality (91), which shall be done as follows.

Proof of the inequality (91). Throughout this part, we denote $a=a(x)$ and $b=b(x)$ for simplicity if there is no confusion. Direct calculation of the derivative for the expression (89) gives

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left\{g(\Theta,Z,W;x)\right\}\Big|_{x=0} = \mathrm{(I)}+\mathrm{(II)},$$

where

$$\mathrm{(I)} \coloneqq \mathbb{E}\left\{-\frac{Z}{2\tau_t^\star}\left[\eta(\Theta+\tau_t^\star Z;\alpha_{t-1}^\star\tau_t^\star)-\eta(\Theta+\tau_t^\star Z;\alpha_t^\star\tau_t^\star)\right]\eta'(\Theta+aZ+bW;\alpha_t^\star\tau_t^\star)\right\}\Big|_{x=0};$$

$$\mathrm{(II)} \coloneqq \mathbb{E}\left\{-\frac{W}{2b}\left[\eta(\Theta+\tau_t^\star Z;\alpha_{t-1}^\star\tau_t^\star)-\eta(\Theta+\tau_t^\star Z;\alpha_t^\star\tau_t^\star)\right]\eta'(\Theta+aZ+bW;\alpha_t^\star\tau_t^\star)\right\}\Big|_{x=0}.$$

Next we shall control each term respectively. For the first term, invoking the Lipschitz property yields $|\eta'(x;\zeta)|\le 1$ and $|\eta(x;\zeta_1)-\eta(x;\zeta_2)|\le|\zeta_1-\zeta_2|$, which further give

$$\mathrm{(I)} \le \mathbb{E}\left[\left|\frac{Z}{2}\right|\cdot|\alpha_{t-1}^\star-\alpha_t^\star|\right] \le |\alpha_{t-1}^\star-\alpha_t^\star|. \tag{93}$$

It remains to study the second term (II), for which the main difficulty lies in the fact that $b(x)\to 0$ as $x\to 0$. It is easy to see that the limit value of (II) is equal to the limiting value of (IIa) + (IIb), where

$$\begin{aligned}\mathrm{(IIa)} &\coloneqq -\frac{1}{2b}\mathbb{E}\left\{W\left[\eta(\Theta+\tau_t^\star Z;\alpha_{t-1}^\star\tau_t^\star)-\eta(\Theta+\tau_t^\star Z;\alpha_t^\star\tau_t^\star)\right]\mathbf{1}(\Theta+aZ+bW\ge\alpha_t^\star\tau_t^\star)\right\};\\ \mathrm{(IIb)} &\coloneqq -\frac{1}{2b}\mathbb{E}\left\{W\left[\eta(\Theta+\tau_t^\star Z;\alpha_{t-1}^\star\tau_t^\star)-\eta(\Theta+\tau_t^\star Z;\alpha_t^\star\tau_t^\star)\right]\mathbf{1}(\Theta+aZ+bW\le-\alpha_t^\star\tau_t^\star)\right\}.\end{aligned}$$

The analysis of the two parts are quite similar, so we only only discuss the first part. Denote

$$\mu(x;Z,\Theta,W) \coloneqq \left[\eta(\Theta+\tau_t^\star Z;\alpha_{t-1}^\star\tau_t^\star)-\eta(\Theta+\tau_t^\star Z;\alpha_t^\star\tau_t^\star)\right]\mathbf{1}(\Theta+a(x)Z+b(x)W\ge\alpha_t^\star\tau_t^\star), \tag{94}$$

To establish expression (91), it suffices to show that

$$\lim_{x\to 0}\left|\frac{1}{2b(x)}\mathbb{E}[W\mu(x;Z,\Theta,W)]\right| \le \left|\alpha_t^\star - \alpha_{t-1}^\star\right|. \tag{95}$$

The remaining part of the current section is then devoted to the proof of (95). Towards this, we make the key observation that $\mathbb{E}[\mu(x;Z,\Theta,W)\mid W=w]$ is a Lipschitz function in w. In order to see this, first notice that the convoluted density of $p_{aZ}*p_\Theta$ is bounded over all $a\in[\tau_t^\star/2,\tau_t^\star]$. Indeed, we can calculate the convolution density by

$$(p_{aZ}*p_\Theta)(z) = \int_{-\infty}^{\infty}\frac{1}{a\sqrt{2\pi}}\exp\left\{-\frac{(\Theta-z)^2}{2a^2}\right\}\mathrm{d}P_\Theta \le \frac{1}{a\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}P_\Theta = \frac{1}{a\sqrt{2\pi}} \le \frac{1}{\tau_t^\star},\quad \forall z\in\mathbb{R}. \tag{96}$$

Next, for any $x\in(0,{\tau_t^\star}^2)$ and $w_1<w_2$, direct calculations yield

$$\begin{aligned}
&\left|\mathbb{E}[\mu(x;Z,\Theta,w_1)] - \mathbb{E}[\mu(x;Z,\Theta,w_2)]\right| \\
&= \left|\mathbb{E}\big[(\eta(\Theta+\tau_t^\star Z;\alpha_{t-1}^\star\tau_t^\star) - \eta(\Theta+\tau_t^\star Z;\alpha_t^\star\tau_t^\star))\cdot(\mathbf{1}(\Theta+a(x)Z+b(x)w_1\ge\alpha_t^\star\tau_t^\star) - \mathbf{1}(\Theta+a(x)Z+b(x)w_2\ge\alpha_t^\star\tau_t^\star)\big]\right| \\
&\le \tau_t^\star|\alpha_t^\star-\alpha_{t-1}^\star|\cdot\mathbb{P}\left\{\alpha_t^\star\tau_t^\star - b(x)w_2 \le \Theta+a(x)Z \le \alpha_t^\star\tau_t^\star - b(x)w_1\right\} \\
&\le b(x)|\alpha_t^\star-\alpha_{t-1}^\star||w_1-w_2|,
\end{aligned}$$

where the last inequality follows from the expression (96) and the fact that $a(x)\in[\tau_t^\star/2,\tau^\star]$ when $0<x<{\tau_t^\star}^2$.

Now we proceed to control $\mathbb{E}[W\mu(x;Z,\Theta,W)]$. First we can write

$$\begin{aligned}
|\mathbb{E}\left[W\mu(x;Z,\Theta,W) - W\mu(x;Z,\Theta,0)\right]| &= \mathbb{E}\left\{W\mathbb{E}\left[\mu(x;Z,\Theta,W)-\mu(x;Z,\Theta,0)|W\right]\right\} \\
&\le b(x)|\alpha_t^\star-\alpha_{t-1}^\star|\mathbb{E}[W^2] \\
&= b(x)|\alpha_t^\star-\alpha_{t-1}^\star|.
\end{aligned}$$

Additionally, by symmetry, we know that

$$\mathbb{E}\left[W\mu(x;Z,\Theta,0)\right] = 0,$$

since $W\sim\mathcal{N}(0,1)$ is independent from $\mu(x;Z,\Theta,0)$. Putting everything together, we conclude that

$$\left|\frac{1}{2b(x)}\mathbb{E}[W\mu(x;Z,\Theta,W)]\right| \le \frac{1}{2}|\alpha_t^\star-\alpha_{t-1}^\star|,$$

thus concluding the proof of the inequality (95). Similarly, one can control (IIb). Taking these collectively with (93), we validated the inequality (91).

D Proofs about the risk curve

D.1 Proof of Lemma 1 and Lemma 2

In this section, we present the proofs of Lemma 1 and Lemma 2 by analyzing each term inside the derivative (37). The proofs are divided into several parts: first in Section D.1.1, we derive the explicit expressions of $F_1(\nu,\delta,\alpha)$, $F_2(\nu,\delta,\alpha)$ and their partial derivatives, which serve as the basis for subsequent analyses. Their limiting values are computed in Section D.1.2 for the case when $\delta\to 1^-$, which in turn establishes Lemma 2.

When it comes to the case $\delta\to 0^+$, we start by stating two crucial lemmas concerning the growth of $\alpha^\star$ and $\nu^\star$, and some partial derivatives, as in Lemma D.1 and D.2. Properties (38b) and (39) are direct consequences of these two lemmas. For property (38a), it turns out that the first-order approximations of both the minuend and subtrahend cancel out. Therefore, we need to resort to the second-order computation of these two terms, which is postponed to Section D.1.3. Putting these together completes the proof of Lemma 1. Finally, the proofs of auxiliary Lemma D.1 and D.2 are deferred to the end of this section.

For notational simplicity, throughout this section, we denote $F_1(\nu,\delta,\alpha)$ as F_1, and similarly for F_2 and all other partial derivative functions when the values of (ν,δ,α) are clear from the context.

D.1.1 Expressions of F_1 and F_2

Let us first express F_1 and F_2 as functions of the density and cumulative density functions of the standard Gaussian distribution (denoted as $\phi(\cdot)$ and $\Phi(\cdot)$ respectively). For F_1, it is easily seen that

$$F_1(\nu,\delta,\alpha) = \epsilon\left[\Phi(-\alpha+\sqrt{\delta}\nu)+\Phi(-\alpha-\sqrt{\delta}\nu)\right]+2(1-\epsilon)\Phi(-\alpha)-\delta. \tag{97}$$

When it comes to F_2, invoking the equality (124a) in Lemma E.1, we arrive at the following decomposition

$$F_2 = F_{21}+\epsilon\delta^{-1}F_{22}+(1-\epsilon)\delta^{-1}F_{23}, \tag{98}$$

where

$$\begin{aligned}
F_{21}(\nu,\delta,\alpha) \coloneqq& \frac{\nu^2}{M^2}-1;\\
F_{22}(\nu,\delta,\alpha) \coloneqq& \int_{-\infty}^{\infty}(\eta(\sqrt{\delta}\nu+z,\alpha)-\sqrt{\delta}\nu)^2\phi(z)\mathrm{d}z\\
=& (\sqrt{\delta}\nu-\alpha)\phi(\alpha+\sqrt{\delta}\nu)+(-\sqrt{\delta}\nu-\alpha)\phi(\alpha-\sqrt{\delta}\nu)\\
&+(\alpha^2+1-\delta\nu^2)\left[\Phi(-\alpha-\sqrt{\delta}\nu)+\Phi(-\alpha+\sqrt{\delta}\nu)\right]+\delta\nu^2;\\
F_{23}(\nu,\delta,\alpha) \coloneqq& \int_{-\infty}^{\infty}\eta^2(z,\alpha)\phi(z)\mathrm{d}z = 2\left[-\alpha\phi(\alpha)+(\alpha^2+1)\Phi(-\alpha)\right].
\end{aligned} \tag{99}$$

Direct computation of the partial derivatives yield the following expressions.

Partial derivatives of F_1. For F_1, we have the following three partial derivatives with respect to α,δ and ν respectively, where

$$\begin{aligned}
\nabla_\alpha F_1(\nu,\delta,\alpha) &= -\epsilon\left[\phi(\alpha-\sqrt{\delta}\nu)+\phi(\alpha+\sqrt{\delta}\nu)\right]-2(1-\epsilon)\phi(\alpha);\\
\nabla_\delta F_1(\nu,\delta,\alpha) &= \epsilon\frac{\nu}{2\sqrt{\delta}}\left[\phi(\alpha-\sqrt{\delta}\nu)-\phi(\alpha+\sqrt{\delta}\nu)\right]-1;\\
\nabla_\nu F_1(\nu,\delta,\alpha) &= \epsilon\sqrt{\delta}\left[\phi(\alpha-\sqrt{\delta}\nu)-\phi(\alpha+\sqrt{\delta}\nu)\right].
\end{aligned} \tag{100}$$

Partial derivatives of F_2. With respect to α, from the decomposition (98), one obtains

$$\nabla_\alpha F_2 = \epsilon\delta^{-1}\nabla_\alpha F_{22}+(1-\epsilon)\delta^{-1}\nabla_\alpha F_{23}, \tag{101}$$

where

$$\begin{aligned}
\nabla_\alpha F_{22}(\nu,\delta,\alpha) =& -2\left[\phi(\alpha+\sqrt{\delta}\nu)+\phi(\alpha-\sqrt{\delta}\nu)\right]+2\alpha\left[\Phi(-\alpha-\sqrt{\delta}\nu)+\Phi(-\alpha+\sqrt{\delta}\nu)\right];\\
\nabla_\alpha F_{23}(\nu,\delta,\alpha) =& -4\left[\phi(\alpha)-\alpha\Phi(-\alpha)\right].
\end{aligned}$$

With respect to δ, a little algebra leads to

$$\begin{aligned}
\nabla_\delta F_2(\nu,\delta,\alpha) &= -\delta^{-2}\left[\epsilon F_{22}+(1-\epsilon)F_{23}\right]+\epsilon\delta^{-1}\nabla_\delta F_{22}\\
&= -\epsilon\delta^{-2}\left[F_{22}-\delta\nabla_\delta F_{22}\right]-(1-\epsilon)\delta^{-2}F_{23}.
\end{aligned} \tag{102}$$

We then further evaluate the right-hand side of the above equation. Recognizing that

$$\nabla_\delta F_{22}(\nu,\delta,\alpha) = \nu^2\left[\Phi(\alpha-\sqrt{\delta}\nu)-\Phi(-\alpha-\sqrt{\delta}\nu)\right],$$

we can guarantee that

$$F_{22}-\delta\nabla_\delta F_{22} = -(\alpha-\sqrt{\delta}\nu)\phi(\alpha+\sqrt{\delta}\nu)-(\alpha+\sqrt{\delta}\nu)\phi(\alpha-\sqrt{\delta}\nu)+(\alpha^2+1)\left[\Phi(-\alpha-\sqrt{\delta}\nu)+\Phi(-\alpha+\sqrt{\delta}\nu)\right].$$

Plugging this into the expression (102) yields the final expression of $\nabla_\delta F_2$. Finally, with respect to ν, it is straightforward to verify that

$$\nabla_\nu F_2(\nu, \delta, \alpha) = 2M^{-2}\nu + \epsilon\delta^{-1}\nabla_\nu F_{22},$$

where

$$\nabla_\nu F_{22}(\nu, \delta, \alpha) = 2\nu\delta\left[\Phi(\alpha - \sqrt{\delta}\nu) - \Phi(-\alpha - \sqrt{\delta}\nu)\right].$$

D.1.2 The limit of $\delta \to 1^-$: proof of Lemma 2

Equipped with these close-form expressions, we are ready to study the the behaviors of $\alpha^\star$ and $\nu^\star$ when $\delta \to 1^-$, and check the limiting orders of all terms in the expression (37).

Limiting orders of $\alpha^\star$ and $\nu^\star$. We first make the observation that for any given $\nu, \alpha > 0$, $F_1(\nu, \delta, \alpha)$ is strictly decreasing in α and increasing in ν. Also we note that $\nu^\star < M$ as $M/\nu^\star = \tau^\star > 1$. Then it is easy to check that

$$0 = F_1(\nu^\star, \delta, \alpha^\star) < F_1(M, \delta, \alpha^\star) < F_1(M, \delta, 0) = 1 - \delta.$$

Taking $\delta \to 1^-$, one can deduce that $F_1(M, \delta, \alpha^\star) \to 0$, which further indicates that $\alpha^\star \to 0$. Now putting $\alpha^\star \to 0$ together with the expression (99), we reach

$$F_{22}(\nu^\star, \delta, \alpha^\star) \to 1; \quad F_{23}(\nu^\star, \delta, \alpha^\star) \to 1 \quad \Longrightarrow \quad F_2(\nu^\star, \delta, \alpha^\star) \to \frac{\nu^{\star 2}}{M^2}.$$

Combining this with the fact that $F_2(\nu^\star, \delta, \alpha^\star) = 0$ ensures $\nu^\star \to 0$.

Limiting order of $\nu^{\star\prime}(\delta)$. With the limiting values of $\alpha^\star$ and $\nu^\star$ in place, we are ready to check the limiting orders of all the terms in expression (37). First, taking the expressions of F_1, F_2 and their derivatives in Section D.1.1 collectively with some algebra, we can guarantee that

$$\begin{aligned}
&\nabla_\alpha F_1 \sim -2\phi(0); \quad \nabla_\delta F_1 \sim -1; \quad \nabla_\nu F_1 \sim 2\epsilon\delta\nu^\star\alpha^\star\phi(0); \\
&\nabla_\alpha F_2 \sim (-4 + 2\epsilon)\phi(0); \quad \nabla_\delta F_2 \sim -1; \quad \nabla_\nu F_2 \sim 2M^{-2}\nu^\star,
\end{aligned}$$

when $\delta \to 1^-$ and all the partial derivatives are evaluated at the point $(\nu^\star, \delta, \alpha^\star)$. Substituting these relations into (37) yields

$$\begin{aligned}
\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1 &\sim -2(1 - \epsilon)\phi(0); \\
\nabla_\nu F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\nu F_1 &\sim -4M^{-2}\phi(0)\nu^\star.
\end{aligned}$$

When ϵ is bounded away from 1 and when M bounded away from ∞, one can easily see that $\nu^{\star\prime}(\delta) \to -\infty$ as $\delta \to 1^-$ (since $0 < \nu^\star \to 0$). We thus conclude the proof of Lemma 2.

D.1.3 The limit of $\delta \to 0^+$: proof of Lemma 1

Before embarking on the main proof, we make note of the following two lemmas concerned with the growth of $\alpha^\star$ and $\nu^\star$ and the derivatives of F_1 and F_2 when $\delta \to 0^+$ which shall be used multiple times. Their proofs of these lemmas can be found in Section D.2 and D.3 respectively.

Lemma D.1. *When $\delta \to 0^+$, the following properties are satisfied*

$$\lim_{\delta\to 0^+} \alpha^\star = +\infty; \quad \lim_{\delta\to 0^+} \sqrt{\delta}\alpha^\star = 0; \quad \lim_{\delta\to 0^+} \frac{\delta\alpha^\star}{\phi(\alpha^\star)} = 2; \quad \lim_{\delta\to 0^+} \nu^\star = \nu_0. \tag{103}$$

Lemma D.2. *When $\delta \to 0^+$, the limiting order of the partial derivatives of F_1 and F_2 are characterized as following:*

$$\lim_{\delta\to 0^+} \frac{\nabla_\alpha F_1}{\phi(\alpha^\star)} = -2; \quad \lim_{\delta\to 0^+} \nabla_\delta F_1 = -1; \quad \lim_{\delta\to 0^+} \frac{\nabla_\nu F_1}{\phi^2(\alpha^\star)} = 4\epsilon\nu_0; \tag{104}$$

$$\lim_{\delta\to 0^+} \alpha^\star \nabla_\alpha F_2 = -2; \quad \lim_{\delta\to 0^+} \alpha^\star\phi(\alpha^\star)\nabla_\delta F_2 = -1; \quad \lim_{\delta\to 0^+} \nabla_\nu F_2 = \frac{2}{\nu_0}, \tag{105}$$

where all the partial derivatives of F_1 and F_2 are evaluated at the point $(\nu^\star, \delta, \alpha^\star)$.

First note that relation (39) in Lemma 1 directly comes from Lemma D.1; Lemma D.2 combined with a little algebra leads to relation (38b). Therefore, to prove Lemma 1, it is only left for us to establish the limiting order of (38a), which shall be done as follows.

Proof of relation (38a). By virtue of Lemma D.2, one immediately notices that the leading terms in $\nabla_\delta F_2 \nabla_\alpha F_1$ and $\nabla_\alpha F_2 \nabla_\delta F_1$ cancel out with each other. Therefore, to characterize the limiting order of $\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1$, it requires us to investigate the second-order terms. Throughout this part, all the partial derivatives of F_1 and F_2 are calculated at the point $(\nu^\star, \delta, \alpha^\star)$ unless otherwise noted.

Second-order terms of $\nabla_\alpha F_1$ and $\nabla_\delta F_1$.

Let us consider $\nabla_\alpha F_1$ and $\nabla_\delta F_1$. We first make the observation that

$$\begin{aligned}\frac{\nabla_\alpha F_1 + 2\phi(\alpha^\star)}{\phi(\alpha^\star)} &= -\epsilon\left[e^{-\delta\nu^{\star 2}/2}\left(e^{\alpha^\star\sqrt{\delta}\nu^\star} + e^{-\alpha^\star\sqrt{\delta}\nu^\star}\right) - 2\right] \\ &\overset{\text{(i)}}{\sim} -\epsilon\alpha^{\star 2}\delta\nu^{\star 2} \overset{\text{(ii)}}{\sim} -2\epsilon\nu_0^2\alpha^\star\phi(\alpha^\star),\end{aligned} \tag{106}$$

where (ii) follows as a direct consequence of (103). Regarding the relation (i), by virtue of the Taylor expansion, we obtain

$$e^{\alpha^\star\sqrt{\delta}\nu^\star} + e^{-\alpha^\star\sqrt{\delta}\nu^\star} = 2 + \delta\nu^{\star 2}\alpha^{\star 2} + o(\delta\nu^{\star 2}\alpha^{\star 2}), \quad e^{-\delta\nu^{\star 2}/2} = 1 - \delta\nu^{\star 2}/2 + o(\delta\nu^{\star 2}).$$

As already shown in Lemma D.1, $\alpha^\star \to \infty$, and therefore, the $\delta\nu^{\star 2}/2$ term is of smaller order compared with $\delta\nu^{\star 2}\alpha^{\star 2}$, which in turn justifies step (i). Here, recall that these limits are taken with respect to $\delta \to 0^+$.

Moreover, direct computation gives

$$\frac{\phi(\alpha^\star - \sqrt{\delta}\nu^\star) - \phi(\alpha^\star + \sqrt{\delta}\nu^\star)}{\sqrt{\delta}\alpha^\star\nu^\star\phi(\alpha^\star)} = \exp\left\{-\delta\nu^{\star 2}/2\right\}\frac{\exp\{\sqrt{\delta}\alpha^\star\nu^\star\} - \exp\{-\sqrt{\delta}\alpha^\star\nu^\star\}}{\sqrt{\delta}\alpha^\star\nu^\star} \sim 2, \quad \delta \to 0^+, \tag{107}$$

where the the last relation is given by L'Hôpital's rule, combined with the facts $\sqrt{\delta}\alpha^\star \to 0$ and $\nu^\star \to \nu_0$ as $\delta \to 0^+$. Taking this collective with Lemma D.1, one arrives at

$$\nabla_\delta F_1 + 1 = \epsilon\frac{\nu^\star}{2\sqrt{\delta}}\left[\phi(\alpha^\star - \sqrt{\delta}\nu^\star) - \phi(\alpha^\star + \sqrt{\delta}\nu^\star)\right] \sim \epsilon\nu_0^2\alpha^\star\phi(\alpha^\star). \tag{108}$$

Second-order term of $\nabla_\delta F_2$.

Moving on to the quantity $\nabla_\delta F_2$, we can rearrange terms to derive the following decomposition

$$\frac{\nabla_\delta F_2 + (\alpha^\star\phi(\alpha^\star))^{-1}}{(\alpha^\star\phi(\alpha^\star))^{-1}} = \underbrace{\frac{\nabla_\delta F_2 + \delta^{-2}F_{23}}{(\alpha^\star\phi(\alpha^\star))^{-1}}}_{=:\Delta_1} - \underbrace{\frac{\delta^{-2}F_{23} - (\alpha^\star\phi(\alpha^\star))^{-1}}{(\alpha^\star\phi(\alpha^\star))^{-1}}}_{=:\Delta_2}. \tag{109}$$

With this decomposition in mind, we proceed to control the two terms above separately.

- **Step 1: Bounding the term Δ_1.** Armed with the expressions of $\nabla_\delta F_2$ and F_{23} in closed-form, one can re-arrange terms and obtain

$$\nabla_\delta F_2 + \delta^{-2} F_{23} = -\epsilon\delta^{-2}\left[f(s) + f(-s) - 2f(0)\right], \tag{110}$$

where $f(s) \coloneqq -(\alpha^\star + s)\phi(\alpha^\star - s) + (\alpha^{\star 2} + 1)\Phi(-\alpha^\star + s)$, for $s = \sqrt{\delta}\nu^\star$. To facilitate analysis of the expression (110), we make note of the following two facts:

  - For every $k \geq 1$, the rescaled derivative $f^{(k)}(s)/\phi(\alpha^\star - s)$ is a polynomial of s and $\alpha^\star$;
  - Since $\alpha^\star \to \infty$ and $\nu^\star \to \nu_0$ as $\delta \to 0^+$, $s = \sqrt{\delta}\nu^\star$ is therefore negligible compared to any polynomial of $\alpha^\star$.

  Leveraging the aforementioned results, one can see that in the Taylor expansion of f around $s = 0$, the non-zero term with the lowest order of s is the dominant term. By further calculating $f^{(2)}(0) = 0$ and $f^{(4)}(0) = 6\alpha^\star\phi(\alpha^\star)$, we see that

$$\nabla_\delta F_2 + \delta^{-2} F_{23} \sim -\epsilon\delta^{-2}\frac{2}{4!}6\alpha^\star\phi(\alpha^\star)(\sqrt{\delta}\nu^\star)^4 \sim -\frac{\epsilon\nu_0^4}{2}\alpha^\star\phi(\alpha^\star).$$

- **Step 2: Bounding the term Δ_2.** Recalling our definition for function F_{23} (cf. (99)), we can express Δ_2 as follows

$$\frac{\delta^{-2}F_{23} - (\alpha^\star\phi(\alpha^\star))^{-1}}{(\alpha^\star\phi(\alpha^\star))^{-1}} = (1 + R_1)(1 + R_2) - 1, \tag{111}$$

where

$$R_1 \coloneqq \frac{\alpha^{\star 3}\left[-\alpha^\star\phi(\alpha^\star) + (\alpha^{\star 2} + 1)\Phi(-\alpha^\star)\right]}{2\phi(\alpha^\star)} - 1; \quad R_2 \coloneqq \left(\frac{2\phi(\alpha^\star)}{\alpha^\star\delta}\right)^2 - 1.$$

Let us consider each term separately. Firstly, directly invoking expression (124d) from Lemma E.1 suggests $R_1 = o(1)$, as $\delta \to 0^+$. To further study the limiting order of R_1, we obtain

$$\begin{aligned}
\lim_{\alpha^\star\to\infty} \alpha^{\star 2} R_1 &= \lim_{\alpha^\star\to\infty} \frac{\alpha^{\star 5}\left[-\alpha^\star\phi(\alpha^\star) + (\alpha^{\star 2} + 1)\Phi(-\alpha^\star)\right] - 2\alpha^{\star 2}\phi(\alpha^\star)}{2\phi(\alpha^\star)} \\
\text{(L'Hôpital's rule)} &= \lim_{\alpha^\star\to\infty} \frac{(2\alpha^{\star 2} - 7\alpha^{\star 4})\phi(\alpha^\star) + (7\alpha^{\star 5} + 5\alpha^{\star 3})\Phi(-\alpha^\star)}{-2\phi(\alpha^\star)} + 2 \\
&= \frac{7}{2}\lim_{\alpha^\star\to\infty} \frac{\alpha^{\star 3}\left[\alpha^\star\phi(\alpha^\star) - (\alpha^{\star 2} + 1)\Phi(-\alpha^\star)\right]}{\phi(\alpha^\star)} - \lim_{\alpha^\star\to\infty} \frac{\alpha^{\star 2}\left[\phi(\alpha^\star) - \alpha^\star\Phi(-\alpha^\star)\right]}{\phi(\alpha^\star)} + 2 \\
&= -6,
\end{aligned}$$

where the last step uses relations (124c) and (124d) from Lemma E.1. This establishes the limiting order $R_1 \sim -6\alpha^{\star -2}$ as $\delta \to 0^+$.

Turning to the term R_2, one can easily conclude from Lemma D.1 that $R_2 = o(1)$ as $\delta \to 0^+$. To further pin down the limiting order of R_2, we recall that

$$\epsilon\left[\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) + \Phi(-\alpha^\star - \sqrt{\delta}\nu^\star)\right] + 2(1 - \epsilon)\Phi(-\alpha^\star) - \delta = 0$$

as $F_1(\nu^\star, \delta, \alpha^\star) = 0$. These allow us to decompose

$$\alpha^{\star 2}\left[\frac{\delta\alpha^\star}{2\phi(\alpha^\star)} - 1\right] = \frac{\epsilon\alpha^{\star 3}[\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) + \Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) - 2\Phi(-\alpha^\star)]}{2\phi(\alpha^\star)} + \frac{\alpha^{\star 2}\left[\alpha^\star\Phi(-\alpha^\star) - \phi(\alpha^\star)\right]}{\phi(\alpha^\star)}. \tag{112}$$

By virtue of the Taylor expansion, we can use similar reasoning as for the expression (110) to arrive at

$$\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) + \Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) - 2\Phi(-\alpha^\star) \sim \alpha^\star \delta \nu_0^2 \phi(\alpha^\star),$$

as $\delta \to 0^+$. Taking this together with the fact that $\delta\alpha^\star \sim 2\phi(\alpha^\star)$ (cf. (103)) reveals that: the first term in the decomposition (112) scales as $\epsilon\nu_0^2\alpha^{\star 2}\phi(\alpha^\star)$. In addition, from the equation (124c), the second term in the decomposition (112) scales as -1 — which is therefore the dominant term as $\alpha^\star \to \infty$. We can therefore conclude that

$$\alpha^{\star 2}\left[\frac{\delta\alpha^\star}{2\phi(\alpha^\star)} - 1\right] \sim -1 \quad \Longrightarrow \quad \frac{2\phi(\alpha^\star)}{\alpha^\star\delta} = 1 + \alpha^{\star -2} + o(\alpha^{\star -2}).$$

As a consequence, we obtain $R_2 \sim 2\alpha^{\star -2}$.

Substituting the limit scalings of R_1 and R_2 into the decomposition (111) yields

$$\frac{\delta^{-2}F_{23} - (\alpha^\star\phi(\alpha^\star))^{-1}}{(\alpha^\star\phi(\alpha^\star))^{-1}} \sim -4\alpha^{\star -2}. \tag{113}$$

Consequently, the above results on Δ_1 and Δ_2 taken collectively with the expression (109) lead to

$$\nabla_\delta F_2 = -(\alpha^\star\phi(\alpha^\star))^{-1}\left[1 - 4\alpha^{\star -2} + o(\alpha^{\star -2})\right]. \tag{114}$$

Second-order term of $\nabla_\alpha F_2$.

We are only left to establish the order of $\nabla_\alpha F_2$. To this end, re-arranging terms in the expression of $\nabla_\alpha F_2$ leads to

$$\frac{\nabla_\alpha F_2 + 2\alpha^{-1}}{\alpha^{-1}} = \underbrace{\frac{\nabla_\alpha F_2 - \delta^{-1}\nabla_\alpha F_{23}}{\alpha^{-1}}}_{=:T_1} + \underbrace{\frac{\delta^{-1}\nabla_\alpha F_{23} + 2\alpha^{-1}}{\alpha^{-1}}}_{=:T_2}.$$

Therefore, it suffices to analyze the limiting order of the two terms T_1 and T_2 separately, which shall be done as follows.

- **Step 1: Bounding the term T_1.** We characterize the leading term of T_1 by examining its Taylor expansion with similar argument as in (110). Specifically, setting $s := \sqrt{\delta}\nu^\star$, we have

  $$\nabla_\alpha F_2 - \delta^{-1}\nabla_\alpha F_{23} = \epsilon\delta^{-1}[f_2(s) + f_2(-s) - 2f_2(0)],$$

  where $f_2(s) := -2\phi(\alpha^\star + s) + 2\alpha^\star\Phi(-\alpha^\star - s)$. It is straightforward to verify that $f_2^{(2)}(0) = 2\phi(\alpha^\star)$, and it follows that

  $$\nabla_\alpha F_2 - \delta^{-1}\nabla_\alpha F_{23} \sim \epsilon\delta^{-1}2\phi(\alpha^\star)(\sqrt{\delta}\nu^\star)^2 \sim 2\epsilon\nu_0^2\phi(\alpha^\star).$$

- **Step 2: Bounding the term T_2.** To calculate the limiting order of T_2, we establish the decomposition

  $$\frac{\delta^{-1}\nabla_\alpha F_{23} + 2\alpha^{\star -1}}{\alpha^{\star -1}} = 2\frac{2\phi(\alpha^\star)}{\delta\alpha^\star}\left[-\frac{\alpha^\star\left[\alpha^\star\phi(\alpha^\star) - (\alpha^{\star 2}+1)\Phi(-\alpha^\star)\right]}{\phi(\alpha^\star)} + \frac{\phi(\alpha^\star) - \alpha^\star\Phi(-\alpha^\star)}{\phi(\alpha^\star)}\right].$$

  As direct consequences of the expressions (124c), (124d) and (103), one can easily see that

  $$\frac{\alpha^\star\left[\alpha^\star\phi(\alpha^\star) - (\alpha^{\star 2}+1)\Phi(-\alpha^\star)\right]}{\phi(\alpha^\star)} \sim -2\alpha^{\star -2}; \quad \frac{\phi(\alpha^\star) - \alpha^\star\Phi(-\alpha^\star)}{\phi(\alpha^\star)} \sim \alpha^{\star -2}; \quad \frac{2\phi(\alpha^\star)}{\delta\alpha^\star} \sim 1,$$

  as $\delta \to 0^+$. Plugging the above relations into the decomposition of T_2 yields $T_2 \sim 6\alpha^{\star -2}$.

In view of the results above on T_1 and T_2, we can conclude that

$$\frac{\nabla_\alpha F_2 + 2\alpha^{\star -1}}{\alpha^{\star -1}} \sim 6\alpha^{\star -2}. \tag{115}$$

Putting all this together. Thus far, we have established the limiting order of $\nabla_\alpha F_1$, $\nabla_\delta F_1$, $\nabla_\delta F_2$ and $\nabla_\alpha F_2$. Combining all the above pieces together, it is easy to justify that

$$\nabla_\delta F_2 \nabla_\alpha F_1 = 2\alpha^{\star -1}\left[1 - 4\alpha^{\star -2} + o(\alpha^{\star -2})\right];$$
$$\nabla_\alpha F_2 \nabla_\delta F_1 = 2\alpha^{\star -1}\left[1 - 3\alpha^{\star -2} + o(\alpha^{\star -2})\right].$$

It immediately suggests that $\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1 = -2\alpha^{\star -3} + o(\alpha^{\star -3})$, which proves the relation (38a), thus completing the proof of Lemma 1.

D.2 Limiting orders of $\alpha^\star$ and $\nu^\star$ when $\delta \to 0^+$: proof of Lemma D.1

In this section, we present the proofs of the four relations in Lemma D.1 in the sequel.

First claim in (103). Recall that F_1 is defined as

$$F_1(\nu, \delta, \alpha) \coloneqq \epsilon \mathbb{P}\left(\left|\nu\sqrt{\delta} + Z\right| > \alpha\right) + (1-\epsilon)\mathbb{P}\left(|Z| > \alpha\right) - \delta.$$

Since the first two terms are non-negative, when $\delta \to 0^+$, setting $F_1(\nu^\star, \delta, \alpha^\star) = 0$ leads to

$$\mathbb{P}\left(\left|\sqrt{\delta}\nu^\star + Z\right| > \alpha^\star\right) \to 0; \quad \mathbb{P}\left(|Z| > \alpha^\star\right) \to 0.$$

From the second expression, it can be immediately concluded that $\lim_{\delta \to 0^+} \alpha^\star = \infty$.

Second claim in (103). In order to study the limiting behavior of $\sqrt{\delta}\alpha^\star$, we first make the observations that $0 < \nu^\star \le M$ as $M/\nu^\star = \tau^\star \ge 1$ (see (15a)) and Φ is a monotonically increasing function such that

$$\Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) < \Phi(-\alpha^\star) < \Phi(-\alpha^\star + \sqrt{\delta}\nu^\star).$$

Consequently, $F_1(\nu^\star, \delta, \alpha^\star)$ can be upper bounded as

$$\begin{aligned} 0 = F_1(\nu^\star, \delta, \alpha^\star) &= \epsilon\left[\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) + \Phi(-\alpha^\star - \sqrt{\delta}\nu^\star)\right] + 2(1-\epsilon)\Phi(-\alpha^\star) - \delta \\ &\le 2\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) - \delta, \end{aligned}$$

which in turn suggests

$$\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) \ge \delta/2.$$

Apply Lemma E.1 with a little algebra to yield

$$2 \ge \frac{\delta}{\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star)} \overset{\text{(i)}}{\sim} \frac{(\alpha^\star - \sqrt{\delta}\nu^\star)\delta}{\phi(-\alpha^\star + \sqrt{\delta}\nu^\star)} \overset{\text{(ii)}}{\sim} \frac{1}{\sqrt{2\pi}}\alpha^\star \delta e^{\alpha^{\star 2}/2},$$

where (i) is a direct consequence of the expression (124b), and (ii) follows from the observations that $\alpha^\star \to \infty$ and $\sqrt{\delta}\nu^\star \le M\sqrt{\delta} \to 0$ as $\delta \to 0^+$. It therefore reveals that $\alpha^{\star 2}\delta = o(1)$, due to the fact that $\alpha^\star \to \infty$. Thus, we complete the proof of the second claim.

Third claim in (103). With the limiting values of $\alpha^\star$ and $\alpha^\star\sqrt{\delta}$ in place, we are now ready to characterize the limiting orders of $\Phi(-\alpha^\star - \sqrt{\delta}\nu^\star)$ and $\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star)$. To this end, we first recognize the following relation

$$\Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) \sim \Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) \sim \Phi(-\alpha^\star) \sim \frac{\phi(\alpha^\star)}{\alpha^\star}, \tag{116}$$

when $\delta \to 0^+$, followed by the property $\alpha^\star\sqrt{\delta}\nu^\star \le M\alpha^\star\sqrt{\delta} = o(1)$. It immediately follows that

$$\begin{aligned} \lim_{\delta \to 0^+} \frac{\delta\alpha^\star}{\phi(\alpha^\star)} &= \lim_{\delta \to 0^+} \frac{\left[\epsilon\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) + \epsilon\Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) + 2(1-\epsilon)\Phi(-\alpha^\star)\right]\alpha^\star}{\phi(\alpha^\star)} \\ &= \epsilon + \epsilon + 2(1-\epsilon) = 2, \end{aligned}$$

where the first equality comes from property $F_1(\nu^\star, \delta, \alpha^\star) = 0$. We thus finish the proof of the third claim.

Fourth claim in (103). In order to understand the limit of $\nu^\star$ as $\delta \to 0^+$, we resort to the decomposition (98) of F_2, where

$$F_2 = \frac{\nu^2}{M^2} - 1 + \epsilon\delta^{-1}F_{22} + (1-\epsilon)\delta^{-1}F_{23}. \tag{117}$$

We claim that it satisfies

$$0 = F_2(\nu^\star, \delta, \alpha^\star) = \frac{\nu^{\star 2}}{M^2} - 1 + \epsilon\nu^{\star 2} + o(1), \quad \text{as} \quad \delta \to 0^+. \tag{118}$$

Taking the above result as given for the moment, one can conclude that $\lim_{\delta\to 0^+}\nu^\star = \nu_0$ as desired, where ν_0 is defined as $\nu_0 \coloneqq M/\tau_0$ and $\tau_0^2 = 1 + \epsilon M^2$ (cf. the expression (35)).

Now it remains to establish the crucial relation (118). To this end, the idea is to characterize the limiting order of each term in the expression (117) in terms of $\nu^\star$, as $\delta \to 0^+$. Let us start with the quantity $\delta^{-1}F_{23}$. Firstly, invoking the equation (124d) from Lemma E.1, we know that $F_{23} \sim 4\alpha^{\star -3}\phi(\alpha^\star)$. As a result, one can write

$$\delta^{-1}F_{23} \sim 4\delta^{-1}\alpha^{\star -3}\phi(\alpha^\star) \sim 2\alpha^{\star -2},$$

where the last relation uses the third claim in (103) that we just proved, namely, $\lim_{\delta\to 0^+}\frac{\delta\alpha^\star}{\phi(\alpha^\star)} = 2$. Since $\alpha^\star \to +\infty$ as $\delta \to 0^+$, we can ensure that $\delta^{-1}F_{23} = o(1)$ as $\delta \to 0^+$.

When it comes to the terms in $\delta^{-1}F_{22}$, we find it useful to define the following function

$$g(x) \coloneqq \phi(x) - x\Phi(-x). \tag{119}$$

One shall then conclude from the expression (124c) that $g(x) \sim x^{-2}\phi(x)$ when $x \to \infty$. With this piece of notation, we can rewrite F_{22} as

$$F_{22}(\nu,\delta,\alpha) = -(\alpha - \sqrt{\delta}\nu)g(\alpha + \sqrt{\delta}\nu) - (\alpha + \sqrt{\delta}\nu)g(\alpha - \sqrt{\delta}\nu) + \Big[\Phi(-\alpha - \sqrt{\delta}\nu) + \Phi(-\alpha + \sqrt{\delta}\nu)\Big] + \delta\nu^2.$$

Let us examine each term on the right-hand side above respectively. For the terms involving g, again applying $\lim_{\delta\to 0^+}\frac{\delta\alpha^\star}{\phi(\alpha^\star)} = 2$ gives

$$\frac{(\alpha^\star - \sqrt{\delta}\nu^\star)g(\alpha^\star + \sqrt{\delta}\nu^\star)}{\delta} \sim \frac{\alpha^{\star -1}\phi(\alpha^\star)}{\delta} \sim \frac{1}{2}; \quad \frac{(\alpha^\star + \sqrt{\delta}\nu^\star)g(\alpha^\star - \sqrt{\delta}\nu^\star)}{\delta} \sim \frac{\alpha^{\star -1}\phi(\alpha^\star)}{\delta} \sim \frac{1}{2}.$$

In addition, for the terms involving the function Φ, by the sandwich relation (116), we arrive at

$$\Phi(-\alpha - \sqrt{\delta}\nu) + \Phi(-\alpha + \sqrt{\delta}\nu) \sim \delta.$$

Plugging the above relations into F_{22} leads to $\delta^{-1}F_{22}(\nu^\star, \delta, \alpha^\star) = o(1) + \nu^{\star 2}$, as $\delta \to 0^+$.

Combining the conclusions about $\delta^{-1}F_{22}$ and $\delta^{-1}F_{23}$, we successfully establish the equation (118), thus finishing the proof of the fourth claim.

D.3 Limiting orders of the partial derivatives: proof of Lemma D.2

We are now positioned to study the limiting orders of the partial derivatives of F_1 and F_2 stated in Section D.1.1. This will be accomplished by taking advantage of Lemma D.1.

Properties concerning F_1 in (104). Recognizing the fact that $\sqrt{\delta}\alpha^\star \to 0$ and $\nu^\star \to \nu_0$ as $\delta \to 0^+$ (as in Lemma D.1), we obtain that $\phi(\alpha^\star - \sqrt{\delta}\nu^\star) \sim \phi(\alpha^\star + \sqrt{\delta}\nu^\star) \sim \phi(\alpha^\star)$. Combining this with the explicit expressions of derivatives of F_1 (cf. (100)), we immediately obtain $\nabla_\alpha F_1 \sim -2\phi(\alpha^\star)$.

Further, taking Lemma D.1 collectively with the relation (107) where

$$\frac{\phi(\alpha^\star - \sqrt{\delta}\nu^\star) - \phi(\alpha^\star + \sqrt{\delta}\nu^\star)}{\sqrt{\delta}\alpha^\star\phi(\alpha^\star)\nu^\star} \sim 2, \quad \delta \to 0^+,$$

we can directly see that

$$\nabla_\nu F_1 \sim 2\epsilon\nu^\star\delta\alpha^\star\phi(\alpha^\star) \sim 4\epsilon\nu_0\phi^2(\alpha^\star)$$
$$\frac{\epsilon\nu^\star}{2\sqrt{\delta}}\left[\phi(\alpha^\star - \sqrt{\delta}\nu^\star) - \phi(\alpha^\star + \sqrt{\delta}\nu^\star)\right] \sim \epsilon\nu_0^2\alpha^\star\phi(\alpha^\star) = o(1)$$

as $\delta \to 0^+$. From the second relation, one can conclude that $\nabla_\delta F_1 \sim -1$, and thus complete the proof of expression (104).

Properties concerning F_2 in (105). Let us turn to the analysis of the partial derivatives related to F_2. In what follows, we shall check the limiting orders of $\nabla_\alpha F_2$, $\nabla_\delta F_2$ and $\nabla_\nu F_2$ respectively, with the assistance of Lemma E.1 and Lemma D.1.

- **Limiting order of $\nabla_\alpha F_2$.** It is useful to recall the decomposition as in expression (101), where

  $$\nabla_\alpha F_2 = \epsilon\delta^{-1}\nabla_\alpha F_{22} + (1-\epsilon)\delta^{-1}\nabla_\alpha F_{23}.$$

  Now we are only left to analyze each term on the right-hand side of the above relation separately. Firstly, the relation (124c) directly yields $\nabla_\alpha F_{23} \sim -4\alpha^{\star -2}\phi(\alpha^\star)$; further recognizing that $\delta \sim 2\alpha^{\star -1}\phi(\alpha^\star)$ (cf. (124b)), we have $\delta^{-1}\nabla_\alpha F_{23} \sim -2\alpha^{\star -1}$.

  To analyze the quantity $\delta^{-1}\nabla_\alpha F_{22}$, we again invoke the definition of function g in (119) to obtain the decomposition

  $$\nabla_\alpha F_{22}(\nu,\delta,\alpha) = -2g(\alpha+\sqrt{\delta}\nu) - 2g(\alpha-\sqrt{\delta}\nu) + 2\sqrt{\delta}\nu\left[\Phi(-\alpha-\sqrt{\delta}\nu) - \Phi(-\alpha+\sqrt{\delta}\nu)\right].$$

  Taking this together with the facts that $g(x) \sim x^{-2}\phi(x)$ when $x \to \infty$, and $\phi(\alpha^\star + \sqrt{\delta}\nu^\star) \sim \phi(\alpha^\star - \sqrt{\delta}\nu^\star) \sim \phi(\alpha^\star)$, leads to

  $$g(\alpha^\star + \sqrt{\delta}\nu^\star) \sim g(\alpha^\star - \sqrt{\delta}\nu^\star) \sim \alpha^{\star -2}\phi(\alpha^\star) \qquad \text{as} \quad \delta \to 0^+.$$

  Finally, we claim that the terms involving Φ are negligible. This can be shown by combining the results in Lemma E.1 with the equality (107). Specifically, one has

  $$2\sqrt{\delta}\nu^\star\left[\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) - \Phi(-\alpha^\star - \sqrt{\delta}\nu^\star)\right] \sim 4\phi(\alpha^\star) = o(1),$$

  where the last step uses $\phi(\alpha^\star) = o(1)$ as $\delta \to 0^+$. Putting these pieces together gives

  $$\delta^{-1}\nabla_\alpha F_{22} \sim -4\delta^{-1}\alpha^{\star -2}\phi(\alpha^\star) \sim -2\alpha^{\star -1}.$$

  As a result, one immediately realizes that $\nabla_\alpha F_2 \sim -2\alpha^{\star -1}$.

- **Limiting order of $\nabla_\delta F_2$.** First recall that $\nabla_\delta F_2 = -\epsilon\delta^{-2}\left[F_{22} - \delta\nabla_\delta F_{22}\right] - (1-\epsilon)\delta^{-2}F_{23}$. As a direct consequence of the relation (124d), one has $F_{23} \sim 4\alpha^{\star -3}\phi(\alpha^\star)$. When it comes to the first term, re-arranging terms in the expression of F_{22} leads to

  $$\begin{aligned} F_{22} - \delta\nabla_\delta F_{22} = &-(\alpha^\star - \sqrt{\delta}\nu^\star)\phi(\alpha^\star + \sqrt{\delta}\nu^\star) - (\alpha^\star + \sqrt{\delta}\nu^\star)\phi(\alpha^\star - \sqrt{\delta}\nu^\star) \\ &+ (\alpha^{\star 2} + 1)\left[\Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) + \Phi(-\alpha^\star + \sqrt{\delta}\nu^\star)\right]. \end{aligned}$$

  Applying (124d) again at $\alpha^\star + \sqrt{\delta}\nu^\star$ and $\alpha^\star - \sqrt{\delta}\nu^\star$ gives

  $$\begin{aligned} (\alpha^\star - \sqrt{\delta}\nu^\star)\phi(\alpha^\star + \sqrt{\delta}\nu^\star) + (\alpha^{\star 2}+1)\Phi(-\alpha^\star - \sqrt{\delta}\nu^\star) &\sim -2\alpha^{\star -3}\phi(\alpha^\star), \\ (\alpha^\star + \sqrt{\delta}\nu^\star)\phi(\alpha^\star - \sqrt{\delta}\nu^\star) + (\alpha^{\star 2}+1)\Phi(-\alpha^\star + \sqrt{\delta}\nu^\star) &\sim -2\alpha^{\star -3}\phi(\alpha^\star), \end{aligned}$$

  which together with $\delta \sim 2\alpha^{\star -1}\phi(\alpha^\star)$ (cf. (124b)) directly validate $\lim_{\delta\to 0^+}\alpha^\star\phi(\alpha^\star)\nabla_\delta F_2 = -1$.

- **Limiting order of $\nabla_\nu F_2$.** It is helpful to recall the expression $\nabla_\nu F_2 = 2M^{-2}\nu^\star + \epsilon\delta^{-1}\nabla_\nu F_{22}$. With Lemma D.1 in place, one has

$$\delta^{-1}\nabla_\nu F_{22} = 2\nu^\star\left[\Phi(\alpha^\star - \sqrt{\delta}\nu^\star) - \Phi(-\alpha^\star - \sqrt{\delta}\nu^\star)\right] \sim 2\nu_0.$$

  As a result, we have $\nabla_\nu F_2 \sim 2(M^{-2}+\epsilon)\nu_0 = 2/\nu_0$, where the last equality follows from the definition of ν_0 as $\nu_0 \coloneqq M/\tau_0$ with $\tau_0^2 = 1+\epsilon M^2$ according to the expression (35).

Thus, we complete the proof of Lemma D.2.

D.4 Proof of Lemma 3

We divide this proof into two parts. In the first part, we characterize the limiting values of $\alpha^\star$ and $\nu^\star/M$ as $\epsilon \to 0$ to establish (42); in the second part, we proceed by calculating the limiting orders of each quantity in the expression (37) for $\nu^{\star\prime}(\delta_0)/M$.

Step 1: limit of $\alpha^\star$ and $\nu^\star/M$ as $\epsilon \to 0$.

- To establish the first statement of the expression (42), we make the observation that: having $F_1(\nu^\star, \delta_0, \alpha^\star) = 0$ yields

$$\epsilon\left[\Phi(-\alpha^\star + \sqrt{\delta_0}\nu^\star) + \Phi(-\alpha^\star - \sqrt{\delta_0}\nu^\star) - 2\Phi(-\alpha^\star)\right] = \delta_0 - 2\Phi(-\alpha^\star),$$

  by recalling the expression of F_1 in (97). It then immediately follows that

$$|\delta_0 - 2\Phi(-\alpha^\star)| \le 4\epsilon.$$

  When $\epsilon \to 0$, we know that $2\Phi(-\alpha^\star) \to \delta_0$, or equivalently, $\alpha^\star \to -\Phi^{-1}(\delta_0/2) =: \alpha_0$.

- In the hope of proving the second statement of the expression (42), we find it helpful to recall the decomposition $F_2 = F_{21} + \epsilon\delta^{-1}F_{22} + (1-\epsilon)\delta^{-1}F_{23}$. From $F_2(\nu^\star, \delta_0, \alpha^\star) = 0$, the following relation holds true

$$\left|\frac{\nu^{\star 2}}{M^2} - 1 + \delta_0^{-1}F_{23}\right| \le \epsilon\delta_0^{-1}\left|F_{22} - F_{23}\right|. \tag{120}$$

  Below we shall demonstrate the fact that $|F_{22} - F_{23}|$ is upper bounded by some constant that only depends on δ_0, which means when $\epsilon \to 0$, the right-hand side of the inequality (120) vanishes to zero. In other words, one can conclude

$$\frac{\nu^\star}{M} \to \sqrt{1 - \delta_0^{-1}F_{23}} \to \sqrt{1 - 2\delta_0^{-1}[-\alpha_0\phi(\alpha_0) + (\alpha_0^2+1)\Phi(-\alpha_0)]},$$

  where the last step follows since $F_{23}(\nu^\star, \delta, \alpha^\star) = 2[-\alpha^\star\phi(\alpha^\star) + (\alpha^{\star 2}+1)\Phi(-\alpha^\star)]$.

  Therefore, it boils down to controlling $|F_{22} - F_{23}|$. From now on, let us consider the scenario when $\alpha^\star < 2\alpha_0$. Recognizing that $\alpha^\star \to \alpha_0 > 0$ as $\epsilon \to 0$, one sees that $\alpha^\star < 2\alpha_0$ holds as long as ϵ is sufficiently small. By virtual of the expression (99), we know that $F_{22}(0, \delta_0, \alpha^\star) = F_{23}(\nu^\star, \delta_0, \alpha^\star)$, and

$$\nabla_\nu F_{22}(\nu, \delta_0, \alpha^\star) = 2\nu\delta_0\left[\Phi(\alpha^\star - \sqrt{\delta_0}\nu) - \Phi(-\alpha^\star - \sqrt{\delta_0}\nu)\right],$$

  which combined with direct calculation gives

$$0 < \nabla_\nu F_{22}(\nu, \delta_0, \alpha^\star) < 2\nu\delta_0\Phi(2\alpha_0 - \sqrt{\delta_0}\nu). \tag{121}$$

  With the help of the above relation, we can further obtain

$$\begin{aligned} 0 \le F_{22} - F_{23} = \int_0^{\nu^\star} \nabla_\nu F_{22}(\nu, \delta_0, \alpha^\star)\mathrm{d}\nu &\le \int_0^{\nu^\star} 2\nu\delta_0\Phi(2\alpha_0 - \sqrt{\delta_0}\nu)\mathrm{d}\nu \\ &= 2\int_0^{\nu^\star\sqrt{\delta_0}} \nu\Phi(2\alpha_0 - \nu)\mathrm{d}\nu \le 2\int_0^{+\infty} \nu\Phi(2\alpha_0 - \nu)\mathrm{d}\nu \le C_{\delta_0}, \end{aligned}$$

  where C_{δ_0} is a constant that only depends on δ_0. Therefore, we complete the proof of the second statement in (42).

Step 2: analysis of quantities in (37). Equipped with the limiting values $\nu^\star$ and $\alpha^\star$, we are ready to analyze those terms that appear in the expression (37). Akin to the previous part, we also assume without loss of generality that $\alpha^\star < 2\alpha_0$ throughout this part.

- **Limiting order of the numerator.** Let us first consider the numerator $\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1$, where the partial derivatives are evaluated at $(\nu^\star, \delta, \alpha^\star)$. With $|\phi(x)| \le 1$ and $\nu^\star \le M$ in mind, by virtue of the equations (100), we can easily verify that

$$|\nabla_\delta F_1(\nu, \delta_0, \alpha) + 1| \le \frac{M}{\sqrt{\delta_0}}\epsilon = \sqrt{\frac{\mathrm{SNR}\cdot\epsilon}{\delta_0}}; \quad |\nabla_\alpha F_1(\nu, \delta_0, \alpha) + 2\phi(\alpha)| \le 4\epsilon$$

with SNR defined in equation (12), which reveals that $\nabla_\delta F_1 \sim -1$ and $\nabla_\alpha F_1 \sim -2\phi(\alpha^\star)$ as $\epsilon \to 0$.

We now turn to the F_2-related quantities. Invoking their explicit expressions as derived in Section D.1.1 yields

$$\left|\nabla_\alpha F_2 - \delta_0^{-1}\nabla_\alpha F_{23}\right| = \epsilon\delta_0^{-1}\left|\nabla_\alpha F_{22} - \nabla_\alpha F_{23}\right|; \tag{122a}$$

$$\left|\nabla_\delta F_2 + \delta_0^{-2} F_{23}\right| = \epsilon\delta_0^{-2}\left|F_{23} - F_{22} + \delta_0\nabla_\delta F_{22}\right|. \tag{122b}$$

We claim that the right-hand sides of both of the above equations vanish as $\epsilon \to 0$. Taking these as given for the moment, we have $\nabla_\alpha F_2 \sim \delta_0^{-1}\nabla_\alpha F_{23}$ and $\nabla_\delta F_2 \sim -\delta_0^{-2} F_{23}$ which further ensure that

$$\begin{aligned}\nabla_\delta F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\delta F_1 &\to 2\delta_0^{-2}\phi(\alpha_0)F_{23} + \delta_0^{-1}\nabla_\alpha F_{23} \\ &\to 2\delta_0^{-2}\phi(\alpha_0)\left[-2\alpha_0\phi(\alpha_0) + (\alpha_0^2+1)\delta_0\right] - 2\delta_0^{-1}\left[2\phi(\alpha_0) - \alpha_0\delta_0\right],\end{aligned}$$

as claimed in the expression (43a). Here, the last step uses $\alpha_0 := -\Phi^{-1}(\delta_0/2)$. For any $0 < \delta_0 < 1$, the limiting value is a negative number, due to the basic relation

$$\Phi(-\alpha_0) \in \left[\phi(\alpha_0)\left(\frac{1}{\alpha_0} - \frac{1}{\alpha_0^3}\right),\ \phi(\alpha_0)\left(\frac{1}{\alpha_0} - \frac{1}{\alpha_0^3} + \frac{1}{\alpha_0^5}\right)\right].$$

Analysis of the expressions (122a) and (122b). Combining $|\phi(x)| \le 1$, $|\Phi(x)| \le 1$ with the expressions of $\nabla_\alpha F_{22}$ and $\nabla_\alpha F_{23}$ in Section D.1.1, one can easily see that

$$|\nabla_\alpha F_{22}| \le 4(1+\alpha^\star) \le 8(1+\alpha_0) \qquad \text{and} \quad |\nabla_\alpha F_{23}| \le 4(1+\alpha^\star) \le 8(1+\alpha_0),$$

given $\alpha^\star < 2\alpha_0$. It is thus clear that (122a) is negligible when $\epsilon \to 0$.

As for the relation (122b), note that we have shown in the previous part that $|F_{23} - F_{22}| \le C_{\delta_0}$, where C_{δ_0} only depends on δ_0. Additionally, it is clear that

$$\begin{aligned}0 \le \nabla_\delta F_{22} &= \nu^{\star 2}\left[\Phi(\alpha^\star - \nu^\star) - \Phi(-\alpha^\star - \nu^\star)\right] \\ &\le \nu^{\star 2}\Phi(2\alpha_0 - \sqrt{\delta_0}\nu^\star) \le \max_{\nu \ge 0}\left\{\nu^2\Phi(2\alpha_0 - \sqrt{\delta_0}\nu)\right\} =: C'_{\delta_0}.\end{aligned}$$

Thus we conclude that $|\nabla_\delta F_{22}| \le C'_{\delta_0}$. Putting the above arguments together leads to the fact that $|F_{23} - F_{22} + \delta_0\nabla_\delta F_{22}|$ is bounded by a universal constant determined by δ_0 and is thus negligible when $\epsilon \to 0$.

- **Limiting order of the denominator.** It remains to study the denominator $M(\nabla_\nu F_2 \nabla_\alpha F_1 - \nabla_\alpha F_2 \nabla_\nu F_1)$. To begin with, from the previous analysis, it is seen that $\nabla_\alpha F_2$ scales as $-2\delta_0^{-1}\left[2\phi(\alpha_0) - \alpha_0\delta_0\right]$ and $\nabla_\alpha F_1$ as $-2\phi(\alpha_0)$ when taking $\epsilon \to 0$. In what follows, we shall analyze $M\nabla_\nu F_2$ and $M\nabla_\nu F_1$ separately. For these two quantities, the following inequalities hold true

$$|M\nabla_\nu F_1| = \sqrt{\mathrm{SNR}\cdot\delta_0\epsilon}\left|\phi(\alpha^\star - \sqrt{\delta_0}\nu^\star) - \phi(\alpha^\star + \sqrt{\delta_0}\nu^\star)\right| \le 2\sqrt{\mathrm{SNR}\cdot\delta_0\epsilon}$$

and

$$\left|M\nabla_\nu F_2 - 2\frac{\nu^\star}{M}\right| = \sqrt{\mathrm{SNR}\cdot\delta_0^{-2}\epsilon}\,|\nabla_\nu F_{22}| = 2\sqrt{\mathrm{SNR}\cdot\epsilon}\nu^\star\left[\Phi(\alpha^\star - \sqrt{\delta_0}\nu^\star) - \Phi(-\alpha^\star - \sqrt{\delta_0}\nu^\star)\right]$$
$$\leq 2\sqrt{\mathrm{SNR}\cdot\epsilon}\nu^\star\Phi(2\alpha_0 - \sqrt{\delta_0}\nu^\star) \leq 2\sqrt{\mathrm{SNR}\cdot\epsilon}C''_{\delta_0},$$

with $C''_{\delta_0} \coloneqq \{\max_{\nu\geq 0}\nu\Phi(2\alpha_0 - \sqrt{\delta_0}\nu)\}$. Taken these collectively, as $\epsilon \to 0$, we achieve

$$M(\nabla_\nu F_2\nabla_\alpha F_1 - \nabla_\alpha F_2\nabla_\nu F_1) \to -4\phi(\alpha_0)\frac{\nu^\star}{M} \to -4\phi(\alpha_0)\sqrt{1 - 2\delta_0^{-1}[-\alpha_0\phi(\alpha_0) + (\alpha_0^2+1)\Phi(-\alpha_0)]},$$

where the last step relies on the relation (42).

Summary. Substituting the above parts on the numerator and the denominator into the expression (37), we conclude that when $\epsilon \to 0$, one has

$$\lim_{\epsilon\to 0}\frac{\nu^{\star\prime}(\delta_0)}{M} < 0,$$

which further indicates that the solution of $\nu^\star/M$ decreases with δ near $\forall\delta_0 \in (0,1)$, as long as ϵ is below a certain threshold.

E Auxiliary lemmas and details

E.1 An example satisfying Assumption 1

Example 1. Let $\{\lambda_t\}$ be a piece-wise constant sequence with

$$\lambda_t = \mu_k, \qquad \text{for } S_{k-1}+1 \leq t \leq S_k, \tag{123}$$

where the length of each piece $s_k \coloneqq S_k - S_{k-1}$ and S_0 is set to be 0. Further choose $\mu_k = 1/\max\{\log k, 1\}$ and $\Lambda_{S_k} = \sum_{i=1}^{S_k}\lambda_i = k^3$, $k \geq 1$.

Proof. Let us now verify that this sequence satisfies Assumption 1. It is straightforward to validate the other inequalities, so we only present the proof for the second relation of (25). In this case, when $S_k \leq t \leq S_{k+1}-1$, direct calculations yield

$$l_t = \sum_{s=1}^{t}|\lambda_s - \lambda_{s+1}|\exp\{-c[\Lambda_t - \Lambda_s]\} = \sum_{j=1}^{k}|\mu_j - \mu_{j+1}|\exp\left\{-c[\Lambda_t - \Lambda_{S_j}]\right\}$$
$$\leq \sum_{j=1}^{k}|\mu_j - \mu_{j+1}|\exp\left\{-ck^3 + cj^3\right\}\exp\left\{-c(t-S_k)\mu_k\right\}.$$

Further, it is easy to verify that $|\mu_j - \mu_{j+1}| \sim 1/(j\log^2 j)$, and as a result,

$$\sqrt{l_t} \lesssim \exp\left\{-\frac{c\mu_k(t-S_k)}{2}\right\}\exp\left\{-\frac{ck^3}{2}\right\}\sqrt{\sum_{j=1}^{k}\frac{1}{j\log^2 j}\exp\{cj^3\}} \overset{(*)}{\lesssim} \exp\left\{-\frac{c\mu_k(t-S_k)}{2}\right\}\frac{1}{k^{3/2}\log k},$$

where $(*)$ follows from the fact that

$$\int_1^k \frac{1}{x\log^2 x}\exp(cx^3)\mathrm{d}x \lesssim \frac{1}{k^3\log^2 k}\exp(ck^3).$$

Finally, we arrive at

$$\sum_{t=1}^{+\infty}\sqrt{l_t} \leq \sum_{k=1}^{+\infty}\left[\sum_{t=S_k}^{S_{k+1}-1}\exp\left\{-\frac{c\mu_k(t-S_k)}{2}\right\}\right]\frac{1}{k^{3/2}\log k} \lesssim \sum_{k=1}^{+\infty}\frac{1}{1-\exp(-c/2\mu_k)}\frac{1}{k^{3/2}\log k} \lesssim \sum_{k=1}^{+\infty}\frac{1}{k^{3/2}} < \infty.$$

□

E.2 Auxiliary lemma for Gaussian distributions

We collect some useful expressions about the standard Gaussian distribution, which shall be used multiple times in the proof of Section D.

Lemma E.1. *The density function and cumulative density function $\phi(\cdot)$ and $\Phi(\cdot)$ of the standard Gaussian distribution obey the following relations:*

$$\int_b^\infty (z-a)^2\phi(z)\mathrm{d}z = (b-2a)\phi(b) + (a^2+1)[1-\Phi(b)]; \tag{124a}$$

$$\lim_{\alpha\to+\infty} \frac{\alpha\Phi(-\alpha)}{\phi(\alpha)} = 1; \tag{124b}$$

$$\lim_{\alpha\to+\infty} \frac{\alpha^2\left[\phi(\alpha)-\alpha\Phi(-\alpha)\right]}{\phi(\alpha)} = 1; \tag{124c}$$

$$\lim_{\alpha\to+\infty} \frac{\alpha^3\left[\alpha\phi(\alpha)-(\alpha^2+1)\Phi(-\alpha)\right]}{\phi(\alpha)} = -2. \tag{124d}$$

Proof of Lemma E.1. To verify the first expression, direct calculations yield

$$\begin{aligned}
\int_b^\infty (z-a)^2\phi(z)\mathrm{d}z &= \int_b^\infty z^2\phi(z)\mathrm{d}z - 2a\int_b^\infty z\phi(z)\mathrm{d}z + a^2\int_b^\infty \phi(z)\mathrm{d}z \\
&= -\int_b^\infty z\mathrm{d}\phi(z) + 2a\int_b^\infty \mathrm{d}\phi(z) + a^2\int_b^\infty \mathrm{d}\Phi(z) \\
&= -\left[-b\phi(b) - \int_b^\infty \phi(z)\mathrm{d}z\right] - 2a\phi(b) + a^2\left[1-\Phi(b)\right] \\
&= (b-2a)\phi(b) + (a^2+1)[1-\Phi(b)].
\end{aligned}$$

The last three equations can be verified similarly by use of L'Hôpital's rule; as an illustration, we provide the proof of the last equation here. By taking the derivatives of both the numerator and the denominator and using a little algebra, we obtain

$$\begin{aligned}
\lim_{\alpha\to+\infty} \frac{\alpha^3\left[\alpha\phi(\alpha)-(\alpha^2+1)\Phi(-\alpha)\right]}{\phi(\alpha)} &= \lim_{\alpha\to+\infty} \frac{4\alpha^3\phi(\alpha) - \alpha^5\phi(\alpha) + (\alpha^5+\alpha^3)\phi(\alpha) - (5\alpha^4+3\alpha^2)\Phi(-\alpha)}{-\alpha\phi(\alpha)} \\
&= \lim_{\alpha\to+\infty} \frac{-5\alpha^2\phi(\alpha) + (5\alpha^3+3\alpha)\Phi(-\alpha)}{\phi(\alpha)} \\
&= 3\lim_{\alpha\to+\infty} \frac{\alpha\Phi(-\alpha)}{\phi(\alpha)} - 5\lim_{\alpha\to+\infty} \frac{\alpha^2\left[\phi(\alpha)-\alpha\Phi(-\alpha)\right]}{\phi(\alpha)} = -2,
\end{aligned}$$

which validates the relation (124d). Thus we complete the proof. □